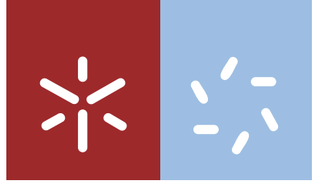

Universidade do Minho
Escola de Ciências

Sara Elisabete Ferreira de Abreu

Estimation of Statistical Cure from Cancer using Population-based Data

outubro de 2015

Sara Elisabete Ferreira de Abreu Estimation of Statistical Cure from Cancer using Population-based Data

UMinho | 2015

Universidade do Minho
Escola de Ciências

Sara Elisabete Ferreira de Abreu

Estimation of Statistical Cure from Cancer using Population-based Data

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efectuado sob a orientação da
Professor Doutor Inês Sousa

e co-orientação do
Professor Doutor Luís Antunes

outubro de 2015

Declaração

Nome: Sara Elisabete Ferreira de Abreu

Endereço electrónico: sarafabreu5@gmail.com

Telefone: 917056598

Número do Bilhete de Identidade: 13739998

Título dissertação:
Estimation of Statistical Cure from Cancer using Population-Based Data

Orientador(es): Professora Doutora Inês Sousa e Doutor Luís Antunes (RORENO)

Ano de conclusão: 2015

Designação do Mestrado: Mestrado em Estatística

1. É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE/TRABALHO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO DO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

Universidade do Minho, _/_/_

Assinatura: _____

"Success is stumbling from failure to failure with no loss of enthusiasm"

Winston S. Churchill

Acknowledgements

I would like to thank my supervisor prof. Inês Sousa for all help she has given me during the study, I am also very grateful for cooperation with doc. Luís Antunes. A special thanks for helpful advices and reading through my draft copies. I would also like to thank the RORENO (North Region of Portugal Cancer Registry) for providing me the data for my work. Additionally, I also highly appreciate the support I have received from my family. The last but surely not least, I am also grateful to my boyfriend for helping me and advising in the use of latex.

Sara Abreu

Abstract

When studying cancer patient survival the interest lies in measure the time until the occurrence of an event, as well as the study of factors associated with the occurrence rates for that event. This is broadly known as survival data and the statistical method for analyzing such data is usually referred to as "survival analysis". In this thesis, standard survival methods, such as the Cox regression model, are recalled and a relative survival approach is preferred over a cause-specific. The analysis is also extended to introduce cure models as a useful tool to analyze and describe cancer survival data. Since some patients actually achieve cure rather than prolong survival, the proportion of patients cured of the disease can be a measure of interest and helpful in monitoring trends in survival of curable disease. Cure models are not widely known and have never been used in North Region of Portugal Cancer Registry (RORENO) reasearches. All these methods are further applied on colorectal and melanoma skin cancer data from RORENO, where this work were developed.

Relative survival estimates and the use of the Cox regression model indicate sex and age as significant prognostic factors in both cancer types studied. With advantage in survival for women and younger patients. Tumour thickness also appeared to be strongly related to the prognosis for melanoma patients with bigger tumour sizes having worst survival rates. Anatomical site where colorectal cancer had developed, colon or rectum, had a different impact in the prognosis of the patient. When developed in the colon the survival estimates were significantly higher. Cure percentages are consistent with survival outcomes, since the identified factors with worse prognosis have also lower cure proportions. According to sex, the proportion of cured patients present a decrease from women (76% for melanoma, 60% for cancer of the colon and 53% for cancer of the rectum) to men (65% for melanoma, 54% for cancer of the colon and 51% for cancer of the rectum). Cure of melanoma and colorectal cancer patients is dependent on age and sex, as well as anatomical site of the cancer cells (colon or rectum) for colorectal cancer patiens.

Efforts in prevention methods are required in order to reduce the percentage of patients presenting with bigger tumour sizes for melanoma. Besides the standard survival analysis to identify prognostic factors and make an overview of the patients survival during the follow-up, it is also important to know how many patients are actually being cured and what factors may have influence on that. This is the only way to understand if efforts and care methods adopted are being successful and actually enable cure from cancer.

Resumo

Na análise de sobrevivência de pacientes com cancro o interesse reside no tempo até a ocorrência de morte, bem como o estudo de fatores com impacto nas taxas de ocorrência para esse evento. Nesta tese, métodos usados na análise de sobrevivência, tais como modelo de regressão de Cox, são recordados e uma abordagem de sobrevivência relativa é preferível ao uso da sobrevivência observada. O estudo teve também como objectivo introduzir os modelos de cura como uma ferramenta útil para analisar e descrever dados de sobrevivência. Uma vez que alguns pacientes alcançam a cura para o cancro em vez de prolongar a sua sobrevivência como paciente com a doença, a proporção de pacientes curados do cancro pode ser uma medida útil e de interesse para compreender tendências dentro dos pacientes que alcançam cura. Os modelos de cura nunca foram utilizados no Registo Oncológico Regional do Norte (RORENO), sendo esta a primeira vez a serem testados com dados de cancro colo-rectal e melanoma disponibilizados pelo RORENO. Estimativas de sobrevivência relativa e modelo de regressão de *Cox* indicam o sexo e a idade como fatores prognósticos significativos em ambos os tipos de cancro estudados, com vantagem na sobrevivência para as mulheres e pacientes mais jovens. Verificou-se também que pacientes com maiores tumores em termos de área e espessura aquando do diagnóstico tendem a ter piores taxas de sobrevivência. O sítio anatómico onde o cancro colo-rectal surgiu e se desenvolveu, cólon ou recto, tem impacto diferente no prognóstico do paciente. Quando desenvolvido no cólon as estimativas de sobrevivência são significativamente mais altas. As percentagens de cura são consistentes com os resultados de sobrevivência, uma vez que os fatores identificados como tendo pior prognóstico também tem as mais baixas proporções de cura. De acordo com o sexo, a proporção de pacientes curados apresenta uma diminuição do sexo feminino (76 % para melanoma, 60 % para o cancro do cólon e 53 % para o cancro do recto) para o masculino (65 % para melanoma, 54% para o cancro do cólon e 51 % para o cancro do recto). Em conclusão, a cura para o melanoma e cancro colo-rectal varia dependendo da idade e sexo do paciente, assim como a localização anatómica das células cancerígenas (cólon ou recto) tem impacto na probabilidade de cura para o cancro colo-retal.
Melhorias nos métodos de prevenção são necessárias a fim de reduzir a percentagem de pacientes que apresentam tamanhos avançados dos tumores de melanoma no diagnóstico. Para além de estudar a sobrevivência dos pacientes durante o período de *follow-up* e os factores com influência no prognóstico, é também relevante conhecer a percentagem de pacientes que alcançam a cura efectiva para o seu cancro e quais os factores possivelmente associados. Só deste modo podemos compreender se os esforços e métodos adoptados no tratamento estão a ser bem sucedidos e permitem alcançar a cura para o cancro.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is a class of diseases characterized for uncontrolled cell growth and tumor formation. There are over 100 types of cancers known with some being a lot more common than others. Nowadays, this disease is affecting millions of people worldwide and it is the main cause of death in economically developed countries. The incidence of cancer globally has increased and forecasts indicate significant increases in the future, due to growth and ageing of the global population alongside with adoption of cancer-causing behaviors, like smoking or poor diet [1]. This is a matter of huge concern and it is affecting our society at economical level and in the way we live. Prevention is always better than cure and considerable efforts were made over decades to comprehend the causes. Valuable progresses were achieved in prevention programs, medical care and detection techniques. Life quality, during and after treatment, as well as long survival times for cancer patients have remarkably improved. And moreover, cure from cancer is increasingly becoming an attainable goal, with a growing number of patients overcoming successfully the disease. Researches and statistical analysis play an important role in comprehending cancer burden, risk factors, survival and cure. Cancer epidemology is the science that studies causes and trends of cancer diseases on populations rather than on separate individuals. Understanding how the disease evolves and what factors may have effect on that evolution and origin is the key to develop treaments and find out ways of prevention. Monitoring health care quality has become an integral part of Portuguese Oncology Institute (IPO) where data are collected and analysed. This work aims to perform a biostatistical analysis using population-based data from RORENO (North Region of Portugal Cancer Registry) to draw conclusions about time trends and prognostic factors associated to colonrectal and melanoma skin cancer. Moreover, it also intends to apply a statistical modelling approach, known as Cure Models, to estimate cure proportion.

1.1 Cancer in Portugal and RORENO

Portugal is no exception regarding to what is happening with cancer worldwide. In 2007, 42374 new cancer cases were diagnosed with 13184 of them in the North Region. Which represented a 7.7% gain in relation to 2006. The most common cancers are breast cancer (incidence rate of 101.6/100000), prostate cancer (113.6/100000) and colorectal cancer with (81.6/100000) [2].

The RORENO (North Region of Portugal Cancer Registry) was created in 1988 at the Portuguese Oncology Institute (IPO) branch of OPorto. Since then, all cancer cases in the north region have to be notified to registry. The geographic area covers OPorto, Braga, Viana do Castelo, Vila Real and Bragança districts (representing about 30% of the total Portuguese population), with is a mix of urban and rural inhabitants similar to Portugal as a whole. Apart from collecting information on new cases as completed and updated as possible, RORENO also analyze the data and manage studies and publications of great value to clinicians, policymakers and researchers.

1.2 Objectives of the Thesis

This thesis focuses on the use of a new and complementary tool in biostatistics to measure cure from cancer and identify factors that may have influence on that. Thus, the primary aims of this thesis are:

1. Make a short overview on the survival analysis concepts.
2. Presentation of a mathematical method named Cure Models as a complementary tool in survival analysis to estimate and model cure from cancer.
3. Applying Cure Models on melanoma and colonrectal cancer data from the North Region of Portugal Cancer Registry to model the number of potentially cured patients.
4. Discuss on the gains of using Cure Models for drawing conclusions about cure.

All the survival analysis and Cure Models application was done in STATA software.

Chapter 2

Survival Analysis

All methodologies and studies presented in this thesis are in a survival analysis framework. The purpose of this chapter is to present a brief introduction to survival analysis and some important key issues in survival studies.

Survival analysis is a branch of statistics focused in the study of the time, T, until an event occurs. In cancer patient survival analysis, event is defined as the death due to cancer. Data are collected during a finite period of time and for some subjects the event of interest (death) may not happen during the follow-up time. In these cases, data is censored. The combination of censoring and differential follow-up makes a problem impossible to solve with standard statistical methods.

To distinguish censoring times from times at which the event was observed for each individual i, with observed time t_i, is defined an event indicator, d_i, that takes the value 0 if the time was censored and 1 if the patient has experienced the event.

Time to event, T, is a positive random variable with distribution function *F(t)* and can be described using the survival function *S(t)*, the hazard function *h(t)* and the cumulative hazard function *H(t)*.

$$S(t) = P(T > t) = 1 - F(t) \tag{2.1}$$

S(t) is a non-increasing function over time and can only take values between 0 and 1. Being 1 at time t=*0* and 0 at time t=∞. The hazard function, *h(t)*, is the instantaneous rate at which events occur, given no previous events.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2.2}$$

The cumulative hazard describes the accumulated risk up to time t,

$$H(t) = \int_0^t \mathrm{h(u)d}u \tag{2.3}$$

The relationship among the functions described above can be expressed mathematically.

$$S(t) = exp(-H(t)) = exp(-\int_0^t \mathrm{h(u)d}u) \tag{2.4}$$

$$h(t) = \frac{d}{dt}H(t) = \frac{f(t)}{S(t)} = -\frac{lnS(t)}{dt} = -\frac{S'(t)}{S(t)} \tag{2.5}$$

2.1 Concepts in Survival Analysis

2.1.1 Population-based Data

Data is the basis for any statistical analysis. Its quality and completeness are the key to the relevance of the study, which in epidemiology research is even more valid.

In cancer patient survival studies, population-based cohort is often used rather than individual cancer patient data because it lies more in the description of patient survival in a demographically representative way [3]. Data collection in population-based studies relies in the compilation of all new cancer cases occuring in a well defined population, usually by the geographical region.

There are some advantages in using population-based data: (1) estimation of distributions and prevalence rates of variables in the reference population can be obtained; (2) risk factor distributions measured at baseline in a study with data periodic updates can be compared with distributions in the future cross-sectional samples, so as to assess risk factors trends over time; (3) a representative sample is the ideal way to carry out unbiased evaluations of relations [4] [5].

In this thesis, it is of interest to use population-based cohort to analyze cancer patient survival in a defined population, that is the north region of Portugal.

2.1.2 Relative Survival

Relative survival is the standard method for quantifying cancer patient survival in population-based data. Relative survival, *R(t)*, uses the all-cause deaths and compares the observed survival, *S(t)*, with that which would have been expected if the cancer patients have had the same mortality rates as in the general population, $S^*(t)$.

$$R(t) = \frac{S(t)}{S^*(t)} \tag{2.6}$$

The expected survival (or background mortality rate) is obtained from national mortality statistics stratified by sex, age, calendar year and possibly other covariates. Although this statistics include deaths due to cancer, the impact on the estimated background risks of death can be negligible. There are several approaches to estimate the expected survival, the most common are the Ederer I and II and the Hakulinen methods [6].

The reason for use relative survival rather than cancer-specific survival is that it does not rely on the classification of cause of death, which is unlikely to be available or poorly reported in population-based studies [7] [8].

On the hazard scale, we can write the overall hazard, *h(t)*, as

$$h(t) = h^*(t) + \lambda(t) \tag{2.7}$$

where $h^*(t)$ is the expected hazard (mortality) rate and $\lambda(t)$ the excess hazard (mortality) rate.

The relative survival and the overall hazard can vary by covariates, $\mathbf{z}$, such as sex, age, stage or calendar year of diagnosis. The expected mortality vary by the stratification factors given in the mortality rates, $\mathbf{z}'$, which are usually a subset of $\mathbf{z}$.

$$R(t;\mathbf{z}) = \frac{S(t;\mathbf{z})}{S^*(t;\mathbf{z}')} \tag{2.8}$$

$$h(t;\mathbf{z}) = h^*(t;\mathbf{z}) + \lambda(t;\mathbf{z}) \tag{2.9}$$

2.1.3 Cox Proportional Hazards Model

The Cox Proportional Hazards Model is the most used model for analysis time to event data on medicine [9]. It explores the relationship between the survival of a patient and several explanatory variables. Cox's method is not a fully parametric model since it doesn't assume any particular distribution for the survival times, instead it assumes that the effects of the different variables on survival are constant over time and are additive in a particular scale [10] . The hazard function for a patient indexed with *i* can be written as

$$h(t;\mathbf{z}_i) = h_0(t)e^{\mathbf{z}_i\beta} \tag{2.10}$$

where β is the vector of regression coefficients and $h_0(t)$ is the so-called baseline hazard function. Since the $h_0(t)$ is only time dependent, the hazard ratio between two different observations does not depend on time

$$\frac{h(t;\mathbf{z_i})}{h(t;\mathbf{z_j})} = \frac{h_0(t)e^{\mathbf{z}_i\beta}}{h_0(t)e^{\mathbf{z}_j\beta}} \tag{2.11}$$

To estimate the β coefficients a partial likelihood function is maximized

$$logL(\beta) = \sum_{i,j=1,\ldots,n} \left\{ \mathbf{z_i}\beta - \mathbf{log}\left[\sum_{\mathbf{j}=\mathbf{t_{(j)}}\geq\mathbf{t_{(i)}}} \mathbf{e}^{\mathbf{z_j}\beta} \right] \right\} \tag{2.12}$$

2.1.4 Survival Analysis Limitations

Since survival time is a difference between the date of death and the time of clinical diagnosis, it is sensitive in changes in either of the dates. Increasings in the 5-year relative survival rate can be affected by lead-time bias and be easily misleading. For instance, if new procedures enable ealier diagnosis without increase the potential of cure and postpone death, the date of death is the same as in the late diagnosis but the survival time is bigger. This leads to increases in survival rates, but actually this situation does not represent effective improvements in survival for cancer patients. Five-year relative survival rate is extremely useful when comparing cancer therapies in a randomized trial but its use is inadvisable in comparisons across time (or place) [11] . Cox regression model is only appropriate when the interest relies in the relative effect of a covariate on the hazard rate [12] . So, classical survival analysis methods do not provide direct estimates of the cure fraction, and thus do not elucidate if the survival improvement over time occurs due to actual cure or prolongation in survival time.

In order to evaluate progress against cancer, one must simultaneously interpret trends in incidence, mortality, survival and cure proportion.

Chapter 3

Cure Models

For many types of cancer, the mortality among patients has seen to return to the same level as in the general population with patients no longer experiencing additional or excess mortality compared with the population from which they are drawn. This is assumed to be a definition of cure at a population level, called population or statistical cure, and differs conceptually from "medical cure" for individual patients [13]. In this situation, we observe that *R(t)* reaches a plateau. For example, in Figure 3.1 relative survival reaches 0.4 and does not go below this.

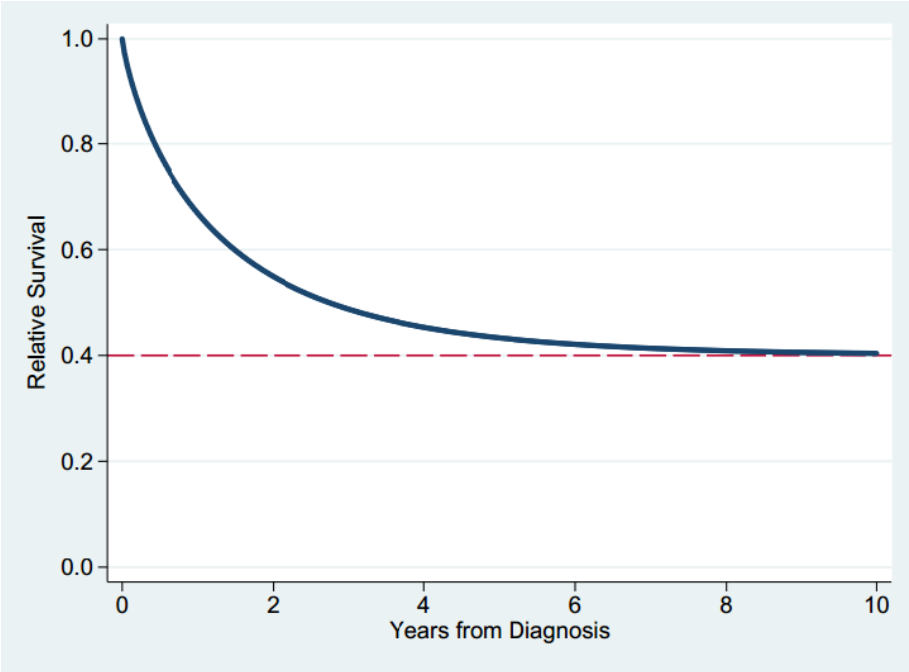

Figure 3.1: Plateau for Relative Survival

Classical survival analysis focuses in the study of time to the event (death) and assume

that every individual is susceptible to the event. Sometimes, however, the data come from a population where a proportion of individuals do not die from the disease and are actually cured. For such cases, reporting the time to event/survival does not bring enough information on the actual condition of the patient and does not take into account the possibility of cure.

Cure models, on the other hand, assume the population of patients divided into two distinct groups: one being cured with respect to the disease and the other one experiencing the event (death from cancer), with separate survival distributions for each. Estimates of the cure fraction can be obtained, which are easier to interpret and more relevant to patients and clinicians than the measures reported from traditional approaches [14]. For the uncured group of individuals measures of the survival time can be provided and, unlike the standard survival methods, are not overestimated because they do not take into account patients who are cured.

Since cure models give information on the cure fraction and survival times of uncured, they elucidate if the survival improvement over time occurs due to effective cure or prolongation in survival time [15], [16]. Another advantage is that they can also provide possibilities for studying temporal trends when analyzing the four possible scenarios.

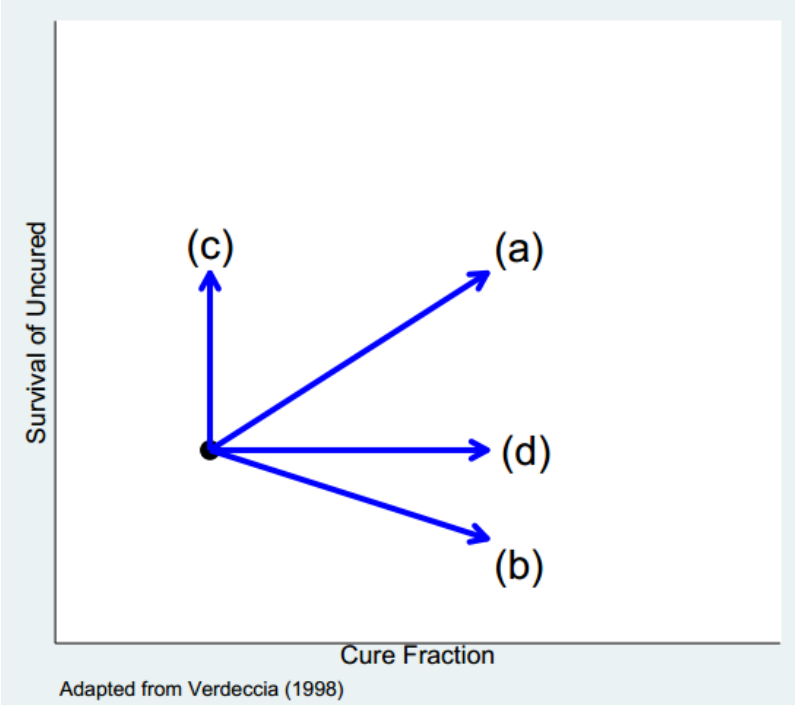

Figure 3.2: Hypothetical changes in the cure fraction and median survival of the uncured between 2 periods of diagnosis

Figure 3.2 shows the different scenarios: (a) represents a general improvement, treatments are able to cure more individuals and provide a longer survival time for those uncured; (b) might happen when more patients are cured and those who did not achieve cure have the worst prognosis. So, the survival of the uncured decrease, while the cure proportion increase; (c) occurs when treatments and procedures only enable to prolong life for cancer patients instead of effective cure. Or, (c) could also occur if we have earlier diagnosis without affecting the time of death. That is, lead time bias and is often suggested as an explanation for improvements in patient survival when using traditional survival methods. An advantage in a cure model framework is that we have information on cure fraction which cannot be affected by lead time bias, and can therefore be used when lead time is a concern; (d) might occur if new procedures able earlier diagnosis leading to greater possibilities for cure. Patients in this case belong to the cure group, thus the cure fraction increase. Still, the survival of the uncured group does not change.

There are, of course, other explanations for these four scenarios and they should be interpreted according to the situation and knowledge of clinical practice.

3.1 Mixture Cure Model

As the name suggests, this model assumes patients as a mixture of two types: ones who experience the event of interest (death from cancer) and the other ones who are cured from the disease, and thus have the same mortality rate as the expected in the general population [17] [18] [19]. The model can be written as

$$S(t) = \pi + (1-\pi)S_u(t) \tag{3.1}$$

where π is the proportion cured and $S_u(t)$ is the survival function of the uncured individuals, for which the parametric Weibull or (more rarely) Lognormal distribution is used. Extending the model to incorporate background mortality, we have

$$S(t) = S^*(t)(\pi + (1-\pi)S_u(t)) \tag{3.2}$$

Which can be expressed on a hazard scale as

$$h(t) = h^*(t) + \frac{(1-\pi)f_u(t)}{\pi + (1-\pi)S_u(t)} \tag{3.3}$$

where $f_u(t)$ is the probability density function associated with $S_u(t)$.

When introducing covariates, $\mathbf{z}$

$$S(t;\mathbf{z}) = S^*(t;\mathbf{z}')(\pi(\mathbf{z}) + (1-\pi(\mathbf{z}))S_u(t;\mathbf{z})) \tag{3.4}$$

Link functions can be used to model the dependence between the covariates and the cure proportion, π;

1. The identity link, $\pi = \mathbf{z}\beta$. Covariate effects are in units of the cure proportion.
2. The logistic link, $log(\frac{\pi}{1-\pi}) = \mathbf{z}\beta$. Covariate effects are expressed as log odds of cure.
3. The $log(-log)$ link, $log(-log(\pi_i)) = \mathbf{z}\beta$. Covariate effects are expressed as (log) excess hazard ratios.

where β is a vector of parameters to be estimated for covariates.

This model allows covariates to be different for cured patients and patients who are not cured. For example, the survival distribution of uncured can depend on only a subset of covariates included to model the cure proportion, or the survival distribution of the uncured can be assumed to not vary by covariates at all.

For survival models, the general log-likelihood can be defined as

$$log\ L = \sum_{i=1}^{N} d_i log\ h(t_i) + log\ S(t_i) \tag{3.5}$$

Where N is the number of individuals and d_i the event indicator, as previously defined.

Estimation is performed by maximum likelihood. The general log-likelihood for the mixture cure model incorporating covariates is

$$logL = \sum_{i=1}^{N} d_i log\left(h^*(t_i;\mathbf{z_i'})\frac{(1-\pi(\mathbf{z_i})) + f_u(t_i;\mathbf{z_i})}{\pi(\mathbf{z_i}) + (1-\pi(\mathbf{z_i}))S_u(t_i;\mathbf{z_i})}\right) + log(\pi(\mathbf{z_i}) + (1-\pi(\mathbf{z_i}))S_u(t_i;\mathbf{z_i})) \tag{3.6}$$

3.2 Non-mixture Cure Model

The Non-mixture model was first introduced to model tumor recurrence, where the cure fraction is the probability that no cancer cells with potential of metastasizing remains [13]. Moreover, this model can also be applied to estimate the cure fraction since it provides an asymptote for the survival function at the cure proportion.

The overall survival function can be written as

$$S(t) = \pi^{F_y(t)} \tag{3.7}$$

When incorporating background mortality, all-cause survival can be expressed as the product of the expected survival and the disease-related (relative) survival

$$S(t) = S^*(t)\pi^{F_y(t)} \tag{3.8}$$

where $F_y(t)$ is a distribution function with $S_y(t)$ its corresponding survival function, and as for the mixture model, a Weibull distribution is often used.

Thus, the overall hazard can be expressed as

$$h(t) = h^*(t) - log(\pi) f_y(t) \tag{3.9}$$

With $f_y(t)$ being the probability density function for $F_y(t)$. Including covariates, $\mathbf{z}$, to model the cure proportion

$$S(t;\mathbf{z}) = S^*(t;\mathbf{z}')exp(log\ \pi(\mathbf{z})(1 - S_y(t;\mathbf{z}))) \tag{3.10}$$

The non-mixture model can be rewritten as a mixture cure model

$$S(t;\mathbf{z}) = S^*(t;\mathbf{z}')\left[\pi(\mathbf{z}) + (1-\pi(\mathbf{z}))(\frac{\pi^{F_y(t;\mathbf{z})} - \pi(\mathbf{z})}{1-\pi(\mathbf{z})})\right] \tag{3.11}$$

The same link functions described for the mixture model can also be used here for modelling of the cure proportion.

The log-likelihood function in this case is

$$log\ L = \sum_{i=1}^{N} d_i log\left(h^*(t_i;\mathbf{z}_i') - log\ \pi(\mathbf{z_i}) f_y(t_i;\mathbf{z_i})\right) + log\left(\pi(\mathbf{z_i}) - log\ \pi(\mathbf{z_i}) S_y(t_i;\mathbf{z_i})\right) \tag{3.12}$$

3.3 Flexible Parametric Cure Model

The models described above assume a Weibull or a Lognormal distribution for the survival function, but these parametric distributions are often inappropriate because of the lack of flexibility to capture the shape of some hazard functions [20], [21]. So, the idea behind the Flexible Parametric Cure Model is to use restricted cubic splines to approximate the shape of the underlying hazard function [22], [23].

Starting from a Weibull survival function, $S(t)$

$$S(t) = exp(-\lambda t^{\gamma}) \tag{3.13}$$

If we transform to the log cumulative hazard scale, we get

$$logH(t) = log(-logS(t)) = log\lambda + \gamma logt \tag{3.14}$$

Which is a linear function of $logt$. Now, rather than assuming linearity with $logt$ restricted cubic splines are used to model the log cumulative hazard function.

$$logH(t) = \gamma_{00} + \gamma_{01}\upsilon_1(x) + ... + \gamma_{0K-1}\upsilon_{K-1}(x) = s(x;\boldsymbol{\gamma_0}) \tag{3.15}$$

Where $x = logt$, K is the number of knots and the j^{th} basis function is defined as

$$\upsilon_j(x) = \begin{cases} x, & for \quad j = 1 \\ (x-k_j)_+^3 - \lambda_j(x-k_1)_+^3 - (1-\lambda_j)(x-k_K)_+^3 & for \quad j = 2,...,K-1 \end{cases} \tag{3.16}$$

In which $u_+ = u$ if $u > 0$ and $u_+ = 0$ if $u \leq 0$, k_1 is the position of the first knot, k_K the position of the last knot, k_j the position of the j^{th} knot and $\lambda_j = (k_K - k_j)/(k_K - k_1)$. The splines used are restricted cubic splines, which has the restriction that the fitted function is forced to be linear before the first knot and after the final knot. The shape of the restricted cubic splines functions is dictated by the available data, so the number and location of knots used has to capture the underlying shape and recent analysis suggest 4-6 knots as being sufficient. With no knots, the spline reduces to a linear function, and these models are equivalent to Weibull, log-logistic and lognormal models.

Introducing covariates, $\mathbf{z}$,

$$logH(t;\mathbf{z}) = s(x;\boldsymbol{\gamma_0}) + \beta\mathbf{z} \tag{3.17}$$

It is of interest to extend this general model idea to a relative survival framework. By integrating equation (2.7), the overall cumulative hazard, $H(t)$, becomes

$$H(t) = H^*(t) + \Lambda(t) \tag{3.18}$$

Where $H^*(t)$ is the cumulative expect hazard and $\Lambda(t)$ is the cumulative excess hazard. The log cumulative excess hazard scale is then modelled using restricted cubic splines to estimate the baseline hazard in the same manner as described above.

$$log(-log\,R(t;\mathbf{z})) = log\Lambda(t;\mathbf{z}) = log\Lambda_0(t) + \beta\mathbf{z} = s(x;\boldsymbol{\gamma_0}) + \beta\mathbf{z} \tag{3.19}$$

In which $\Lambda_0(t)$ is the cumulative baseline function and covariates having an additive effect.

The reasons for modelling on the log cumulative hazard scale instead of the log hazard scale are: the log cumulative hazard is a relatively stable function with easier capture of its shape; it is also easier to transform to the survival and hazard functions without the need for numerical integrations and under the proportional hazards assumptions covariate effects are interpreted as hazard ratios [20].

When cure is reached the excess hazard rate is zero and the cumulative excess hazard will be, therefore, constant after this point in time. Cure proportion estimation, in a flexible parametric model, is done by forcing the log cumulative excess hazard to not only be linear but also to have zero slope after the last knot. All spline functions are zero before the first knot except the linear, $\upsilon_1(x) = x$, and imposing constraints on the parameter, γ_{01}, it is possible to determine the slope of the spline before the first knot. But this is supposed to happen after the last knot, and the only way to get it is by treating the knots in reversed order, "backwards", so that all splines variables except the linear take the value zero after the last knot. Restriction on the parameter for the linear spline variable ($\gamma_{01} = 0$) impose a cure point and enable estimation of the cure proportion. The spline basis, $\upsilon_j(x)$, are then defined

$$\upsilon_j(x) = \begin{cases} x, & for \quad j = 1 \\ (k_{K-j+1} - x)_+^3 - \lambda_j(k_K - x)_+^3 - (1-\lambda_j)(k_1 - x)_+^3, & for \quad j = 2,...,K-1 \end{cases} \tag{3.20}$$

The relative survival function for the flexible parametric cure model becomes

$$R(t) = exp\Big(- exp(\gamma_{00} + 0 \times \upsilon_1(x) + ... + \gamma_{0K-1}\upsilon_{K-1}(x))\Big) \tag{3.21}$$

It can be seen that the flexible parametric cure model is a special case of a non-mixture cure model with $\pi = exp(-exp(\gamma_{00}))$, and $F_y(t) = exp(\gamma_{02}\upsilon_2(x) + ... + \gamma_{0K-1}\upsilon_{K-1}(x))$. $F_y(t)$ is a distribution function as long as the excess mortality is not negative, which is very uncommon. As for the non-mixture and mixture model, the flexible parametric cure model can be written as a proportional hazards model, as long as no time-dependent effects are modelled.

When incorporating covariates and interactions between covariates and splines for time,

$$R(t) = exp\Big(- exp(\gamma_{00} + \beta \mathbf{z})\Big) exp\Big(\gamma_{02}\upsilon_2(x) + ... + \gamma_{0K-1}\upsilon_{K-1}(x) + \sum_{i=1}^{D} s(x;\gamma_{\mathbf{i}})\mathbf{z_i}\Big) \tag{3.22}$$

Where D is the number of time-dependent covariate effects and $s(x;\gamma_{\mathbf{i}})$ is the spline function for the i^{th} time-dependent effect. The constant parameters, γ_{00} and β , are used to model the cure proportion and the time-dependent parameters are used to model the distribution function of $F_y(t)$. The constraint of a zero effect for the linear spline term has to be incorporated for each modeled spline function. Since all splines variables take the value zero after the last knot, the constant parameter, γ_{00}, in equation (3.22) , is the log cumulative excess hazard at and beyond the last knot for the reference group, and can therefore be used to predict cure.

All parameters are estimated by maximum likelihood. Cure proportion and the survival of uncured can be predicted in the same way as for the non-mixture model, and the median survival time of uncured, or any other percentile, is predicted using the Newton-Raphson algorithm [24] .

As for the mixture and non-mixture models, the flexible parametric cure model can be written as a non-proportional hazards model, as long as time-dependent effects are modelled. In cancer patient survival studies it is very common to have time-dependent effects with interactions between covariates and splines for time.

Chapter 4

Case Study Melanoma

Melanoma is a type of skin cancer that develops in the pigment cells present in the skin. It has the ability, in later stages, to spread (or metastasize) to other parts of the body and cause serious illness and death. There has been a worldwide increase in incidence of melanoma among all caucasian populations over the last decades [25]. A recent study performed with data from European cancer registries shows that the highest rates were seen for the Scandinavian and north-western European countries, while Portugal and Spain had the lowest incidence rates [26] . Much of the increase in melanoma occurence has been suggested to be caused by the rise in excessive exposure to sunlight alongside with skin type. That explains why Caucasian individuals and people who are susceptible to red sunburn have the highest risk of developing the disease over Africa, Asia and Middle-East individuals. Despite increasing trends in melanoma incidence, the prognosis for those affected has improved. This is most likely due to earlier detection and new adopted treatment care techniques. In fact, many studies report that prognosis is mainly influenced by tumour thickness and ulceration [27] and this leads to the high importance of effective early detection. But other factors such sex, age at diagnosis and anatomical site are also known to affect patients survival [28].

4.1 Data

The study population were all patients with a diagnosis of skin melanoma cancer in the north region of Portugal between 1998 and 2003 with follow-up to the end of 2013. Patients that have been also diagnosed with other cancers, as well as patients with unknown status were excluded, making up a total of 760 eligible individuals for the analysis. The original RORENO registrations has information on a lot of variables, however records from 1998-2003 are extremely incomplete. For that reason the variables considered in

the study had to be the ones suggested in latter studies with the most completeness in the available data:

- **Sex**

  Identifies the patient gender: 1-Male, 2-Female;

- **Age Group**

  The variable age at diagnosis is recoded into categories commonly used (15-44, 45-54, 55-64, 65-74, 75+);

- **Typology**

  Specifies the zone type in which patients live, according to the official classification from INE (Portuguese Institute of Statistics): 1-APR, 2-AMU, 3-APU. A region is classified as predominantly rural area (APR) if does not belong to a locality with a population of 2 000 or more inhabitants, or if has a population density equal or below 100 inhabitants per km^2. Averagely urban (AMU) is an area that complies, at least, one of the following requirements: (1) the greater value of the average between the weight of population and of area in the parish's total is urban space, with the weight of the area in predominantly rural space surpassing 50% of the parish's total area; (2) the greater value of the average between the weight of population and of area in the parish's total is semi-urban space, with the weight of the area in predominantly rural space not surpassing 50% of the parish's total area; (3) the parish is the location of the municipal hall and has a population equal or below 5 000 inhabitants; (4) the parish contains totally or partially a locality with a population of 2 000 or more inhabitants and less than 5 000 inhabitants, with the weight of the population in the parish's total population or in the locality's total population being equal or higher than 50%. Lastly, a region is labeled predominantly urban area (APU) when follow one of these requirements: (1) the greater value of the average between the weight of population and of area in the parish's total is urban space, with the weight of the area in predominantly rural space not surpassing 50% of the parish's total area; (2) the parish is the is the location of the municipal hall and has a population of more than 5 000 inhabitants; (3) the parish contains totally or partially a locality with a population of 5 000 or more inhabitants, with the weight of the population in the parish's total population or in the locality's total population being equal or higher than 50% [29] .

- **T**

  This variable measures the tumour size/thickness into two new categories formed by grouping: T1-T2, T3-T4;

- **Anatomical Site**

  Region of the human body where the tumour is located. Its recoded into four categories: Head/Neck, Trunk, Upper Extremity and Lower Extremity.

Table 4.1: Cohort characteristics

	Frequency (%)	Number of deaths during follow-up	Percentage dying during follow-up
Sex			
Male	285 (37.50%)	155	54.39
Female	475 (62.50%)	199	41.89
Age Group			
15-44	199 (26.18%)	44	22.11
45-54	133 (17.50%)	40	30.08
55-64	140 (18.42%)	52	37.14
65-74	153 (20.13%)	103	67.32
75+	135 (17.76%)	115	85.19
Typology			
Rural	48 (6.32%)	33	68.75
Averagely Urban	102 (13.42%)	49	48.04
Urban	607 (79.87%)	270	44.48
Unknown	3 (0.39%)	2	66.67
T			
T1-T2	395 (51.97%)	165	41.77
T3-T4	109 (14.34%)	75	68.81
Unknown	256 (33.68%)	114	44.53
Anatomical Site			
Head/Neck	144 (18.95%)	80	55.56
Trunk	180 (23.68%)	82	45.56
Upper Extremity	72 (9.47%)	28	38.89
Lower Extremity	216 (28.42%)	100	46.30
Unknown	148 (19.47%)	64	43.24
Total	760 (100%)	354	46.58

Table 4.1 presents simple statistics of the variables. There is not a similar proportion of males and females in the cohort. More women were diagnosed with the disease

comparing to men, comprising 63% of the total cohort. The most common age group at diagnosis was 15-44 years, for the other age groups the number of patients was very similar, around 140 diagnosis between 1998-2003. The percentage dying during follow-up increased from the newest to the oldest group, suggesting age as a risk factor in prognosis. The number of patients who lived in a urban area is enormous, representing 80% of the study population. The other ones are from a moderately urban area and only 6% resided in the countryside. It should be noted that in the region where study population are draw 68% of the people reside in urban areas, while 21% and 11% are from moderately urban and rural areas, respectively.

The variable T is unknown for 34%, because this information wasn't necessarily required in the registries in 1998-2003. Among ones with information on this parameter, there were a lot more beeing diagnosed with a smaller size of the tumour. The number of deaths during follow-up was higher in the group of patients with tumour sizes in the biggest category. The regions of the body were tumours started to appear were diverse. Melanoma on the lower extremity was the most common, followed by trunk and head/neck, while cases on the upper extremity were just a few (9%).

4.2 Survival Estimates

The purpose of this section is to present an overview of the survival for the set of patients. The survival rates were calculated for each variable in order to compare the influence of the diferent groups in the survival. The relative survival method for cancer survival estimation was Ederer II [30] .

4.2.1 Sex

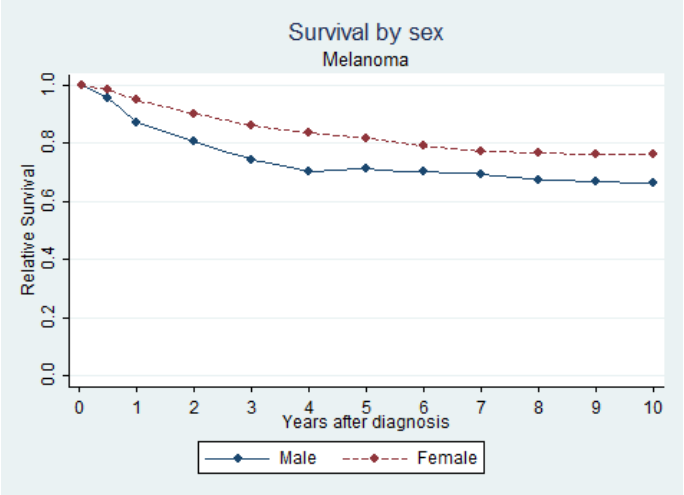

Figure 4.1: Relative Survival by Sex

Table 4.2: Survival by Sex

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
Male	76.5%	60.0%	51.2%	p-value<0.001
Female	87.1%	71.1%	63.9%	
Relative Survival				
Male	80.6%	70.2%	66.3%	
Female	90.2%	79.0%	76.4%	

Figure 4.1 suggests a clear difference in survival among genders. Women with melanoma appear to have better prognosis than men. In Table 4.2, 5-year relative survival is 79% for females and decreases for 70 % in males, while after 10 years of follow-up the survival rate becomes 76% for women and 68% for men. From the first year after diagnosis to the tenth there was a relative survival declined of about 14%for both sexes.

4.2.2 Age Group

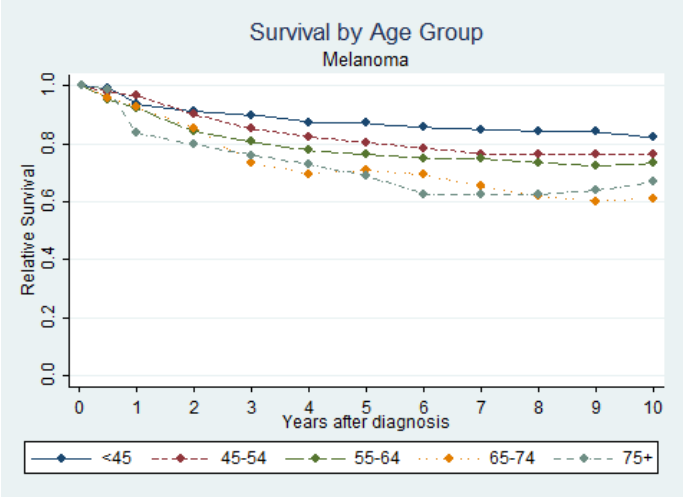

Figure 4.2: Relative Survival by Age Group

Table 4.3: Survival by Age Group

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
15-44	93.5%	86.4%	80.9%	p-value<0.001
45-54	96.2%	79.0%	72.9%	
55-64	91.4%	72.9%	65.7%	
65-74	90.9%	62.8%	45.1%	
75+	76.9%	43.3%	22.4%	
Relative Survival				
15-44	93.6%	87.1%	82.2%	
45-54	96.6%	80.5%	76.2%	
55-64	92.2%	76.3%	73.4%	
65-74	92.8%	70.9%	61.2%	
75+	83.6%	68.9%	66.8%	

Looking at Table 4.3 is easy to see that in the first years after diagnosis survival estimates are more similar to each other and tend to differ over time. During the all follow-up period there is indication of different patterns in survival and the prognosis seems to get worse from the 15-44 to 75+ age group. The relative survival curve of the 75+ age group rises a little bit after 7 year following diagnosis. This is medically impracticable and it may

be due to the use of a inappropriate life table, deficient follow-up (it may be possible that not all deaths were reported) or can also be the case that these patients received better treatments that improved their survival.

4.2.3 Typology

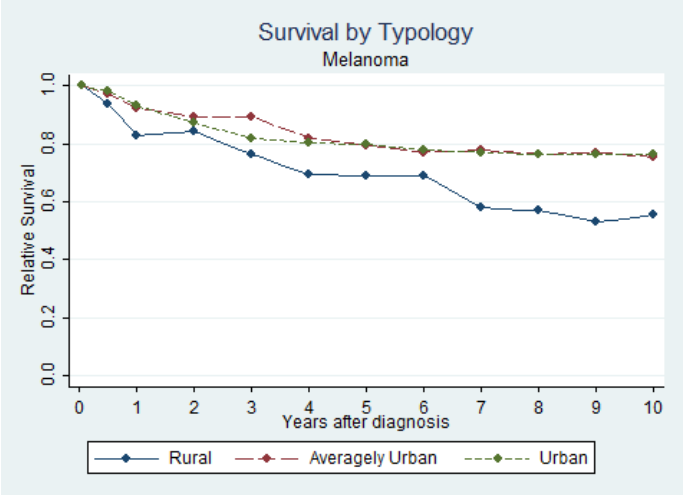

Figure 4.3: Relative Survival by Typology

Table 4.4: Survival by Typology

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
Rural	79.2%	56.3%	37.5%	p-value=0.001
Averagely Urban	90.2%	70.6%	58.8%	
Urban	90.9%	71.5%	61.1%	
Relative Survival				
Rural	82.3%	67.5%	52.7%	
Averagely Urban	92.1%	78.6%	73.2%	
Urban	92.7%	78.6%	74.2%	

Both Figure 4.3 and Table 4.4 shows that survival is very similar among melanoma patients from urban and averagely urban areas, while the patients that live in rural regions have shown lower levels in survival, specially beyond six years after diagnosis.

4.2.4 T

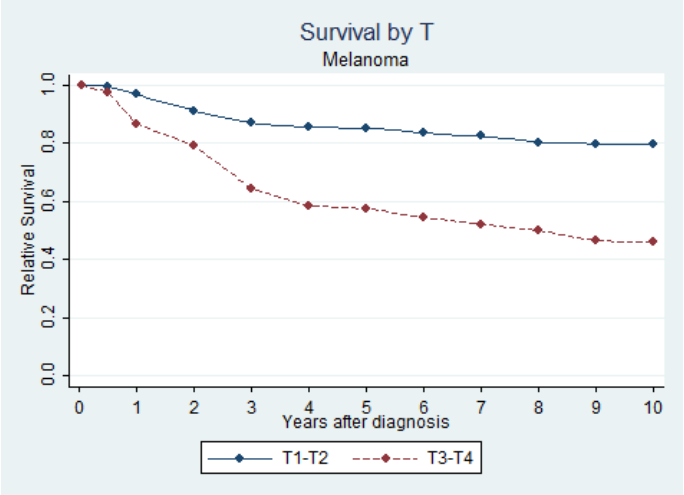

Figure 4.4: Relative Survival by T

Table 4.5: Survival by T

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
T1-T2	94.9%	77.2%	65.1%	p-value<0.001
T3-T4	84.4%	51.4%	37.6%	
Relative Survival				
T1-T2	96.9%	85.2%	79.8%	
T3-T4	86.6%	57.5%	46.1%	

Patients diagnosed with smaller tumours size/thickness showed better chances to resist the disease. Table 4.5 present a survival rate of about 85% in the end of the fifth year. Whereas, the ones with tumours categorized as T3-T4 revealed more difficulty struggling the disease with a survival estimate of 46% in the end of the 10 years of follow-up period.

4.2.5 Anatomical Site

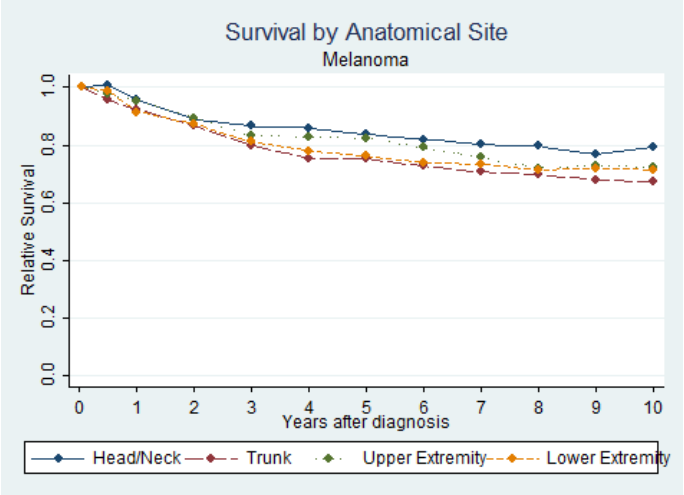

Figure 4.5: Relative Survival by Anatomical Site

Table 4.6: Survival by Anatomical Site

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
Head/Neck	91.7%	68.1%	53.5%	p-value=0.1364
Trunk	91.1%	70.6%	58.3%	
Upper Extremity	94.4%	77.8%	63.9%	
Lower Extremity	89.8%	69.8%	60.5%	
Relative Survival				
Head/Neck	95.6%	81.4%	80.3%	
Trunk	92.3%	75.2%	67.0%	
Upper Extremity	95.4%	82.4%	72.3%	
Lower Extremity	91.4%	76.1%	71.4%	

Relative survival curves in Figure 4.5 and estimates from Table 4.6 indicate that there is no big difference between the prognosis of patients diagnosed with tumours in distint anatomical sites, since the results of the survival estimates do not vary that much.

4.3 Cox regression model for survival

The aim of this section is to apply a Cox regression model for the achievement of factors with influence on the survival time of a patient. A Cox model was built under backward selection method, including at each step the variable with the largest increase in the χ^2 model or equivalently the largest reduction in the log-likelihood. No relevant interactions between the considered variables were identified. As a result, the final model is shown in the next table with estimated hazard ratios associated with the set of covariates, as well as their 95% confidence intervals and corresponding p-values.

Table 4.7: Cox model fitted to available data from Melanoma skin cancer patients

Years after Diagnosis Prognostic Factor	Category	Hazard Ratio	95% CI	p-value
Sex	Male	1		
	Female	0.70	]0.54,0.91[	0.007
Age Group	<65	1		
	65-74	2.83	]2.06,3.87[	<0.001
	75+	4.94	]3.59,6.80[	<0.001
Typology	Rural	1		
	Averagely Urban/Urban	0.63	]0.41,0.98[	0.041
T	T1-T2	1		
	T3-T4	2.18	]1.65,2.88[	<0.001

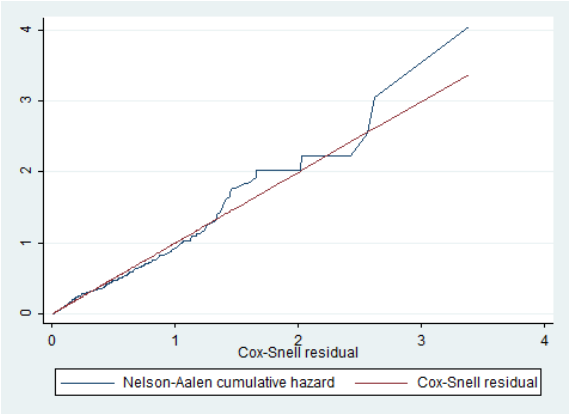

Figure 4.6: Cox-Snell Residual

To evaluate the fit of the model Cox-Snell residuals were plotted. Figure 4.6 mostly confirm the adequacy of the Cox model to fit the data.

The predictor anatomical site is not in the final model, with summary results presented in Table 4.7, because it was clearly not statistically significant, p-value=0.88. Considering age group as a prognostic factor, there was no statistical difference among patients bellow 65 years old. However, from looking at the hazard ratios the model indicates that as a patient change from an age lower than 65 to the age group between 65 to 74 years old (assuming all other variables helding constant) the risk of death increases in 183%. If the age agroup was altered to less than 65 to more than 75 years old, the rate of death increases by 394%. When going from a male to female patient the risk of death decreases in 30%. Patients diagnosed with tumor thickness evaluated as T3-T4 have 118% more risk of dying, comparing with ones diagnosed with T1-T2 and considering all other variables constant. Altough there is no significant statistic difference in prognosis for patients from urban and averagely urban areas, when comparing to rural areas patients there is significance of a decrease in 37% the risk of death.

4.4 Estimating and Modeling cure with Cure Models

In this section the cure models presented in Chapter 3 were applied to the data to demonstrate its applicability and potential to study temporal trends of the cure proportion.

The mixture and non-mixture cure models did not converge, probably because of its known weakness when relative survival is high or inappropriate distributions to describe the survival of the uncured patients. A flexible parametric cure model approach was considered with backward elimination method, starting with all variables as well as interactions between them and using likelihood-ratio tests to test which models fitted the data best. To evaluate how sensitive were the results to different numbers or locations of knots, the most statistically significant model was fitted with different numbers and places of knots. Results confirmed that the models are generally insensitive to that as long as knots are placed over the whole follow-up period and the last knot is positioned at the last observed death time. The final model only includes age group and sex. Estimated relative survival curves of the flexible parametric cure models, stratified by each variable were plotted in Figures 4.7 and 4.8 against life table estimates of relative survival, with 95% confidence intervals, obtained using Ederer II method.

Figure 4.8 shows the flexible parametric fit of observed data by age group and it is obvious that patients above 65 years old fail the assumption of statistical cure needed for cure models application, since its survival curve did not reach a plateau.

For both sexes, age groups 45-54 and 55-64 the flexible parametric cure model give a good fit of observed data, indicating evidence of statistical cure since the survival curves

level off within 10 years of follow-up. For the age group bellow 45 years old the model failed to converge, because the survival was high (above 80%). For patients above age of 65 the graphs confirm that the assumption of cure is not reasonable and, therefore, cure models should not be fitted and interpreted. The next reported results and considearations only take into account the ages proved to be under the cure assumption.

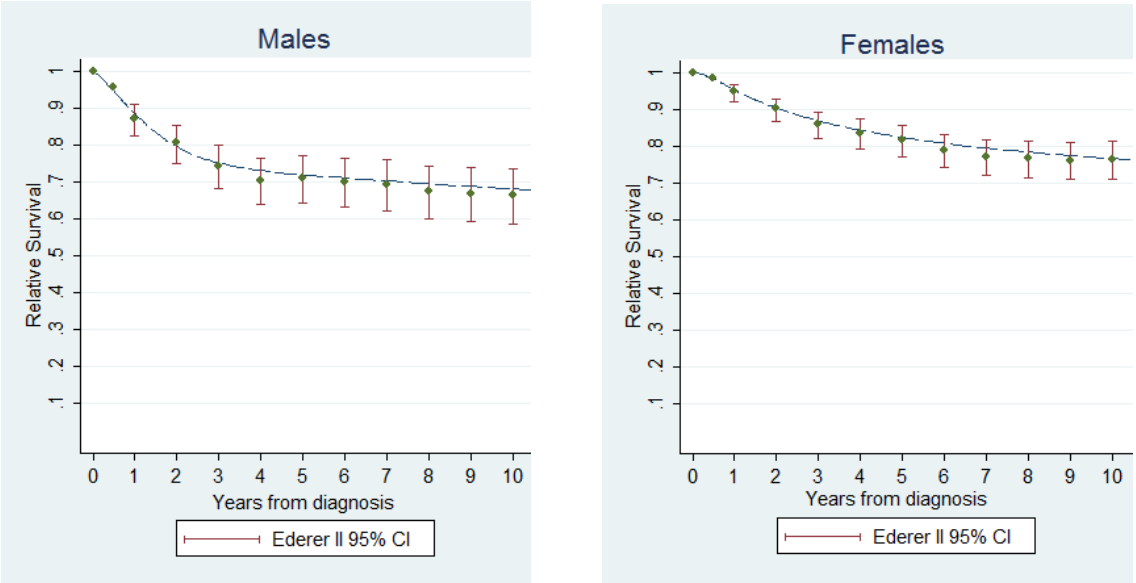

Figure 4.7: Predicted Survival from the Flexible parametric cure model, compared to life table estimates of relative survival for both genders

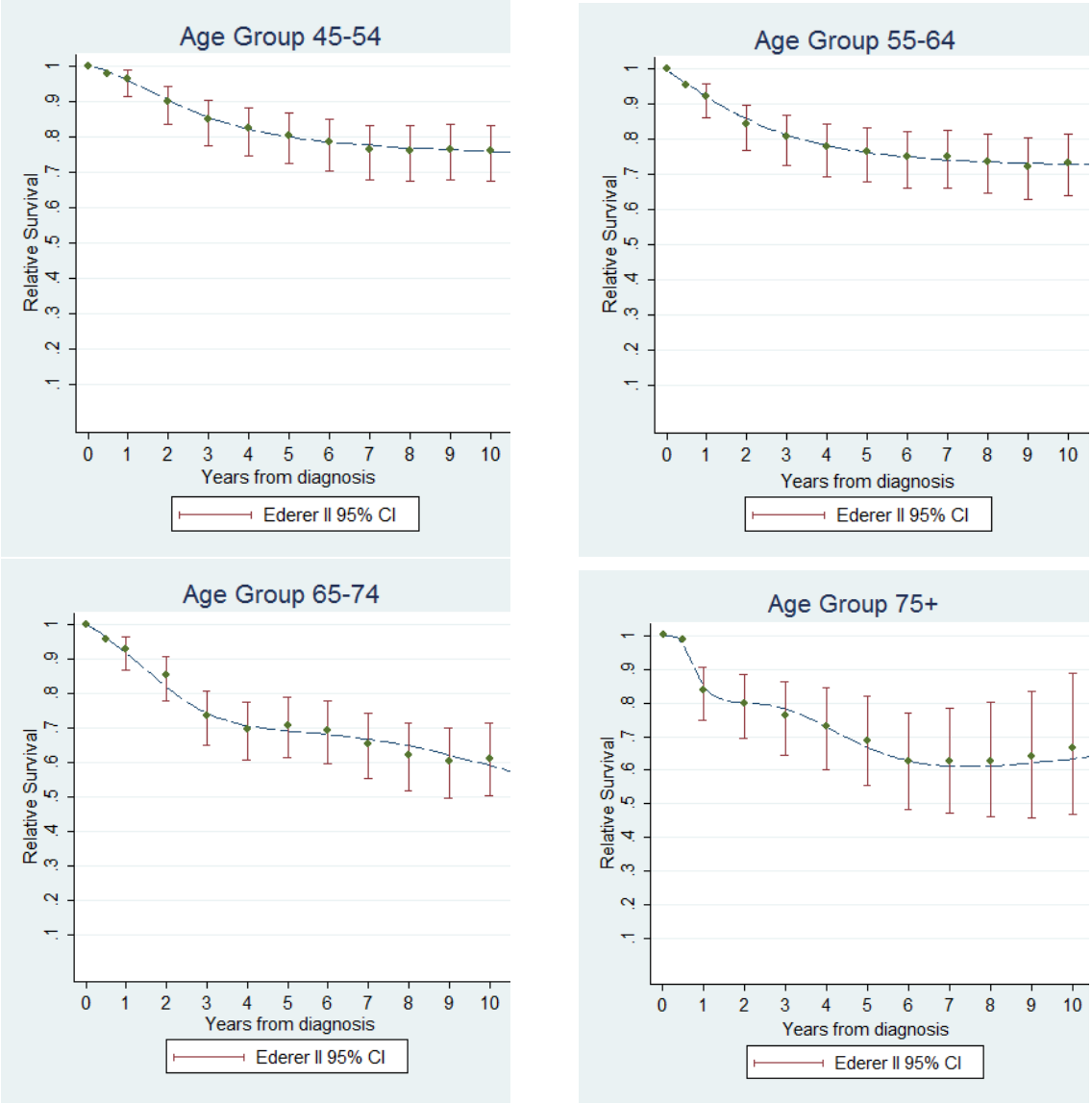

Figure 4.8: Predicted Survival from the Flexible parametric cure model, compared to life table estimates of relative survival for age groups

Table 4.8: Estimated proportion cured (%) and 95% confidence intervals by sex and age group

	Age 45-54	Age 55-64	**All 45-64**
Men	69.07]51.59 , 81.32[	63.19]48.36 , 74.83[	**64.83]54.14 , 73.63[**
Women	79.54]68.37 , 87.13[	71.26]58.04 , 80.98 [	**75.62]67.33 , 82.09[**
All	**75.84]66.66 , 82.82[**	**66.70]58.13 , 73.90[**	

Table 4.8 shows that cure fraction is lower as we go from the newest age group to the oldest for both men and women. Results also vary according to gender, with women

achieving better estimates ranging from 80% in ages between 45 and 54 to 71% in the following ages up to 64. Men's cure percentages do not reach 70%, being 69% among younger patients in the study and 63% for the others. However the uncertainty regarding this estimate is higher, since the confidence interval magnitude for the age group 45-54 lies between 52% and 81%. For the older patients the confidence interval is also wide with the lower limit being below 50%.

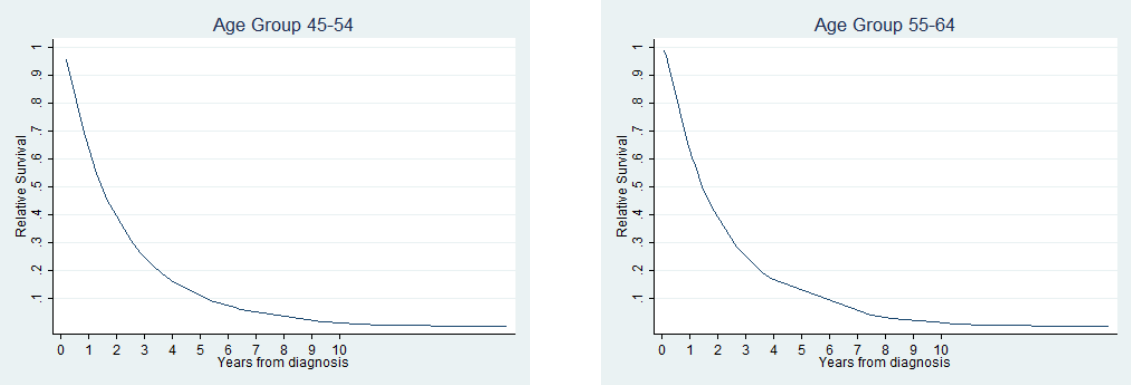

Figure 4.9: Predicted Survival of Uncured for Men

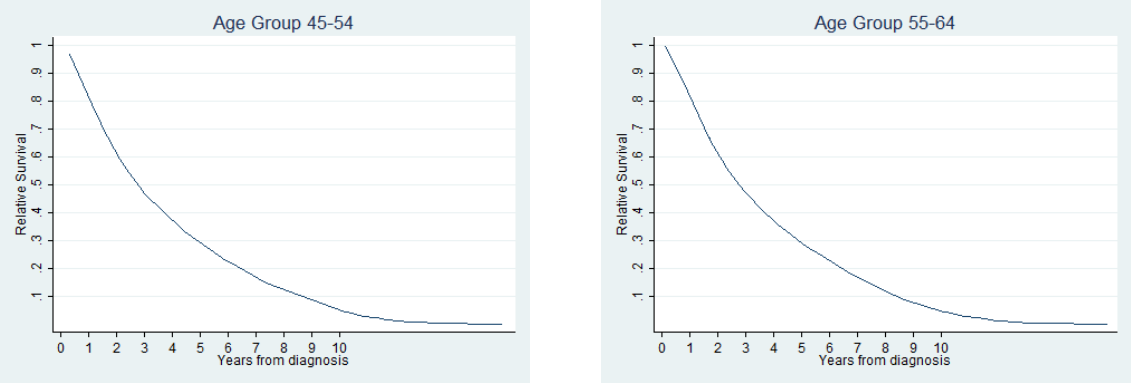

Figure 4.10: Predicted Survival of Uncured for Women

The survival of the uncured in men plotted in Figure 4.9 seem to have a more notable decrease within the 5 years of follow-up. While the group of uncured women, represented in Figure 4.10, appears to resist better the disease.

4.5 Discussion

In this chapter, an exploratory survival analysis was performed to illustrate relative survival behaviour during the 10 years follow-up for melanoma patients from North Region of Portugal Cancer Registry. The aim was to understand how patient features influence

and vary the survival over time. Relative survival curves by gender in Figure 4.1 and estimates in Table 4.2 corroborate what many studies report, that women have a better prognosis. Men have relative survival results below 10% of the ones estimated for women during the whole period. In the first year after being diagnose with melanoma, the relative survival was quite high (81% for men and 90% for women). Then, after 5 years the survival declined 11% for both and from the fifth to the tenth year the survival curve began to plateau. These results support the evidence of long-term survival in melanoma skin cancer. The prognostic potential of age is also not surprising. It was confirmed that survival reduces with advanced age at diagosis and is most pronounced among patients on the age of 75 or above [31] . About the evaluation on survival according to the zone type in which patients live, we can say that there is no difference among ones from averagely urban and urban areas. However, people who live in rural regions had worse survival estimates. Nevertheless, it is important to notice that there are only about 6% of patients from rural areas in the entire cohort which may not be enough reliable. Our findings on tumor thickness influence on survival testify it as an important prognostic predictor. Patients diagnosed with a tumour size/thickness classified as in the most advanced stage had higher mortality. Therefore, early detection remain of high importance in the treatment/long-term survival success. In Section 4.3 a Cox regression model was built up to identify the potential risk factors associated with survival probability. This analysis get us closer to the prognostic profile of a specific patient and may help the clinician both in the communication of risk and in the follow-up strategy. The final model include sex, age group, typology and tumour size/thickness as significant prognostic factors. Age group is the most important prognostic factor as well as size. In fact, patients above 65 years old at diagnosis showed a risk of dying by the disease much bigger than younger ones. And the risk is even worse if we consider patients with 75 or older. The model corroborate what relative survival estimates suggest for tumour size/thickness influence. Actually, being diagnosed with tumour size/thickness in the most advanced stage is twice as likely to compromise the death risk. This information is of great value giving insights in how crucial early detection is. That is why it is so important to raise awareness, strengthening or developing clinical methods and techniques in order to boost vigilance and potential to a detection as early as possible. There is also statistical significance regarding gender distinction in survival trends with a better prognostic for women. Among patients from urban and averagely urban areas there is no significant differences in survival behaviour, but when comparing to patients who come from rural regions the survival trends were not similar. These patients showed higher mortality and a worse prognostic. However, it should be noted that there are only about 6% of patients from rural areas in the entirity cohort which may not be enough representative or may be confounded with differences

in other factors like sex, tumour thickness, anatomical site and so on.

Cure models can be more appropriate and a useful alternative to the standard Cox proportional-hazards regression model when presence of long-term survivors is clear. Since we have a population with long-term survival that verifies the assumption of cure, the interest now lies in knowing the proportion of patients cured of their disease and what happen to those that are not cured. Therefore, additional information can be gained from using a cure model analysis beyond a standard Cox analysis. Measure the cure fraction can be more relevant, easier to interpret and also provide greater possibilities for studying temporal trends. In this study, we were able to estimate the cure percentages by age group and gender. The other covariates available in the data, typology, tumour size/thickness and anatomical site did not fit the observed data to reach a plateau within the follow-up period. Cure proportion estimates are consistent with the findings obtained by cox regression model and relative survival estimates with women and younger age groups having better cure percentages. Younger patients are likely to tolerate treatments better than older ones, and therefore more likely to achieve cure. In the group of uncured patients, men mortality has higher rates with a five year relative survival of about 10%. While women seem to resist for a longer period of time, reaching a survival below 10% only after 9 years of follow-up. Greater cure rates in women indicate biological differences and responses to disease and treatment between genders. Nevertheless, another factors may also have impact like earlier detection or younger ages at diagnosis for women, or even social and psychological features with distinct influence in prognosis for both genders.

Chapter 5

Case Study Colorectal Cancer

Colorectal cancer is the term for the malignant tumours that develop in the large intestine, either the colon or the rectum. These two types of cancer are usually referred as the same, since they presence various features in common and can only differ in terms of treatment procedures. Also known as bowel cancer, this kind of disease has been reported as the second most common cancer in Europe [32] . However, it has been proven not to be uniformly incident throughout the world. Many studies report increases in the incidence among economically transitioning countries, whereas rates stabilized or decreased in long-standing economically developed countries. This reflects the adopting results of western lifestyles and behaviors, like poor diets, physical inactivity and smoking. Mortality rates are also higher in low-resource countries and continue to rise, while in economically developed countries they have declined [33] . This is a matter of huge concern and had led to efforts in developing medical treatment and techniqes that improved the survival and long-term outcome of colorectal patients. Once the cure percentage and median survival among patients is increasing, the factors that may have an influence in the prognosis were also tempted to understand. Age and stage at diagnosis seem to be the main important prognostic factors for this cancer patients [34] . In Portugal, colorectal cancer is the second most common cancer afecting both men and women. In 2008 the incidence rate of colorectal cancer cases among men was (146,4/100.000) and (90,2/100.000) for women [35] .

5.1 Data

The data used for this study consisted in all adults from both sexes above 15 years-old diagnosed with malignant colorectal cancer between 1998 and 2003 in the north region of

Portugal. Information on the patient records was obtained from RORENO registrations comprising a total of 4489 eligible records for the analysis.

- **Sex**

  Identifies the patient gender: 1-Male, 2-Female;

- **Age Group**

  The variable age at diagnosis is recoded into categories commonly used (15-44, 45-54, 55-64, 65-74, 75+);

- **Typology**

  Specifies the zone type in which patients live. Coded in the same way as in melanoma case study.

- **Topography**

  Indicates in which part of the large instestine the cancer developed: 1-Colon, 2-Rectum.

Table 5.1: Cohort characteristics

	Frequency (%)	Number of deaths during follow-up	Percentage dying during follow-up
Sex			
Male	2493 (55.54%)	1674	67.15
Female	1996 (44.46%)	1219	61.07
Age Group			
15-44	246 (5.48%)	106	43.09
45-54	490 (10.92%)	224	45.71
55-64	949 (21.14%)	497	52.37
65-74	1534 (34.17%)	998	65.06
75+	1270 (28.29%)	1068	84.09
Typology			
Rural	448 (9.98%)	317	70.76
Averagely Urban	698 (15.55%)	457	65.47
Urban	3247 (72.33%)	2104	64.80
Unknown	96 (2.14%)	15	15.63
Topography			
Colon	2872 (63.98%)	1826	63.58
Rectum	1617 (36.02%)	1067	65.99
Total	4489 (100%)	2893	64.45

The results in Table 5.1 seem to show different scenarios regarding gender. There was a bigger incidence of this cancer in men than in women and prognosis also seems to be worse for men, since the percentage dying during follow-up was bigger. The most affected ages were over 65 years old representing about 63% of the total cohort. Deaths during treatment were also higher between the older patients, specially the ones above 75 years old.

5.2 Survival Estimates

This section aims to present an overview of the survival for the set of patients. The survival rates were estimated for each variable in order to compare the influence of the diferent groups in the survival. The relative survival method for cancer survival estimation was Ederer II.

5.2.1 Sex

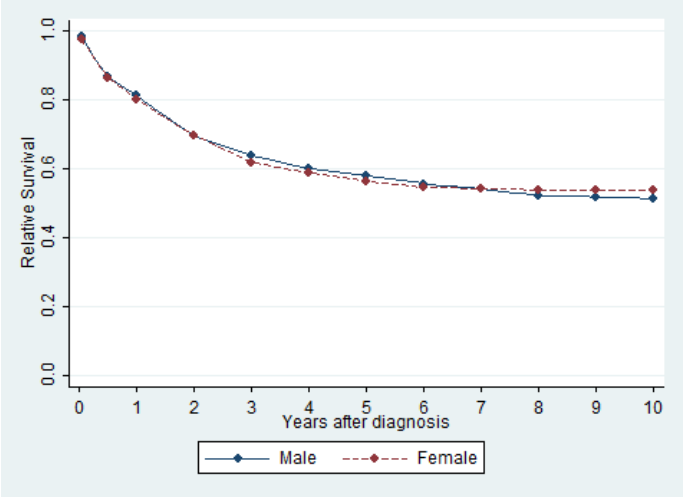

Figure 5.1: Relative Survival by Sex

Table 5.2: Survival by Sex

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
Male	64.6%	44.2%	34.1%	p-value=0.001
Female	65.7%	46.3%	39.9%	
Relative Survival				
Male	69.5%	55.5%	51.2%	
Female	69.4%	54.5%	53.5%	

Observed survival rates seem to indicate a difference in prognosis among genders, with better survival rates for women during the 10 years after diagnosis. When looking to relative survival, the difference is not that clear. Figure 5.1 suggests similar survival for men and women until the first two and a half years, then men seem to resist better to the disease but only until seven years after diagnosis, after that female rates show better results for long term survival.

5.2.2 Age Group

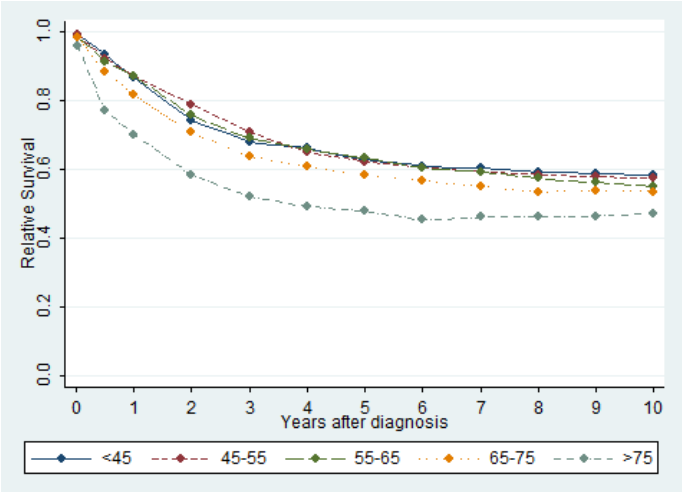

Figure 5.2: Relative Survival by Age Group

Table 5.3: Survival by Age Group

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
15-44	73.7%	60.1%	56.8%	p-value<0.001
45-54	77.9%	58.7%	54.3%	
55-64	74.1%	56.2%	48.2%	
65-74	67.1%	47.5%	37.9%	
75+	49.1%	25.6%	15.7%	
Relative Survival				
15-44	74.0%	60.9%	58.2%	
45-54	78.6%	60.5%	57.5%	
55-64	75.6%	60.4%	55.1%	
65-74	70.7%	57.0%	54.1%	
75+	58.3%	45.2%	47.2%	

Analyzing the survival by age group from Figure 5.2 and Table 5.3 stands out that patients diagnosed with more than 75 years old had the worst chances of beating the disease, followed by the age group bellow with ages between 65 and 75. Persons diagnosed younger than 65 years old reaveled better survival estimates, with long term survival results decreasing as we go along age groups.

5.2.3 Typology

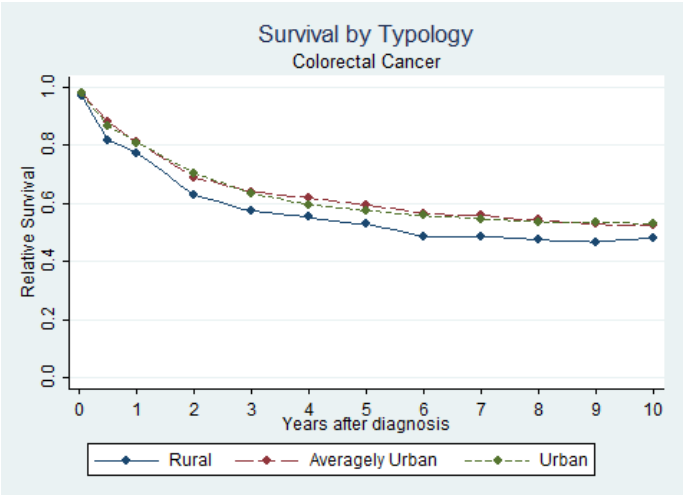

Figure 5.3: Relative Survival by Typology

Table 5.4: Survival by Typology

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
Rural	58.1%	38.6%	31.5%	p-value=0.0213
Averagely Urban	64.8%	46.9%	37.6%	
Urban	66.1%	45.8%	37.3%	
Relative Survival				
Rural	62.8%	48.6%	47.8%	
Averagely Urban	69.0%	56.6%	52.4%	
Urban	70.4%	55.7%	53.0%	

Relative survival curves in Figure 5.3 and estimates in Table 5.4 suggest that patients who live in rural places showed inferior survival rates during the all follow-up period compared with persons who reside in more urbanized regions.

5.2.4 Topography

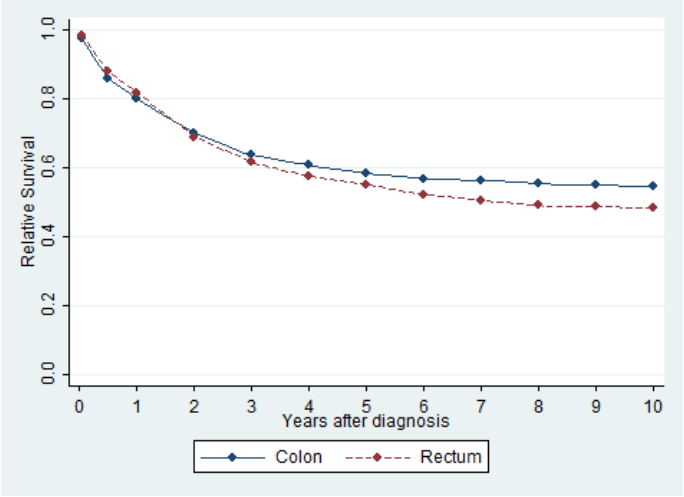

Figure 5.4: Relative Survival by Topography

Table 5.5: Survival by Topography

Years after Diagnosis	1	5	10	Log-Rank Test
Observed Survival				
Colon	65.3%	46.0%	37.7%	p-value=0.1612
Rectum	64.8%	43.6%	35.0%	
Relative Survival				
Colon	69.8%	56.7%	54.6%	
Rectum	68.8%	52.2%	48.2%	

In the first two years after diagnosis survival estimates from Table 5.5 were identical for both types of cancer , being near 70% in the first year. However, after the second year the percentage of cancer in the rectum patients who deceased exceeded the ones with cancer in the colon.

5.3 Cox regression model for survival

Similar to what was presented in previous chapter, in this section a Cox regression model is fitted for the population in study. Backward selection method was also applied and no significant interations between variables were identified.

Table 5.6: Cox model fitted to available data from Colorectal cancer patients

Years after Diagnosis Prognostic Factor	Category	Hazard Ratio	95% CI	p-value
Sex	Male	1		
	Female	0.84	]0.78,0.91[	<0.001
Age Group	<55	1		
	55-64	1.22	]1.05,1.39[	0.008
	65-74	1.69	]1.49,1.91[	<0.001
	75+	3.21	]2.84,3.64[	<0.001
Topography	Colon	1		
	Rectum	1.10	]1.02,1.19[	0.012

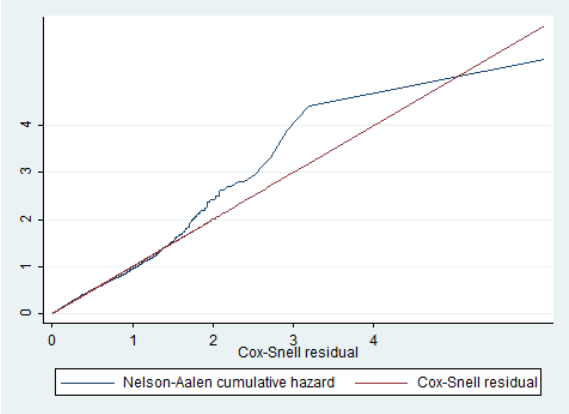

Figure 5.5: Cox-Snell Residual

Figure 5.5 shows that hazard function follows the 45 degree line very closely except for very large values of time, therefore the final model fits data well.

Interpreting hazard ratios from the final model results in table 5.6, we see that in women the rate of relapse is 16% lower than in man (considering all other variables constant). Age groups bellow 55 years old did not revealed significant statistic differences between them, but when comparing to older ages there were actually differences in survival. Assuming age lower than 55 as a reference group, we see that when comparing to a patient in 55-64 age group, the risk of death increases by 21%. If we go to an age between 65 and 74 the risk is 69% higher. And when comparing the reference group to ones with 75 or more years old, the risk has a remarkable rise by 221%. Analyzing predictor topography, if we go to a patient with cancer in the colon to one with cancer developed in the rectum there is indication of 10% more risk of dying.

5.4 Estimating and Modeling cure with Cure Models

A Flexible parametric cure model was applied to all variables in the data, as well as interactions between them. Backward variables elimination by likelihood-ratio tests was also used to get the final model. The number and place of knots sensitivity was evaluated and once again the results confirmed that the model was invariant to that. So, the final model uses the default knots positions. Figures 5.6, 5.7 and 5.8 show the estimated relative survival from Ederer II plotted against predicted survival from the model to visually assess its fit. Figure 5.7 shows that observed data of patients above 75 years old it is not properly fitted by the model, because life table estimates tend to increase in the last years of follow up. And despite its confidence intervals get bigger reflecting uncertainty and possibility of leveling of, patients in these ages will be excluded from the model. For all the age

groups under 75 the assumption of cure is verified with estimates of relative survival tended to level off within 10 years of follow-up. The final parametric cure model include all variables in the study with exception of typology.

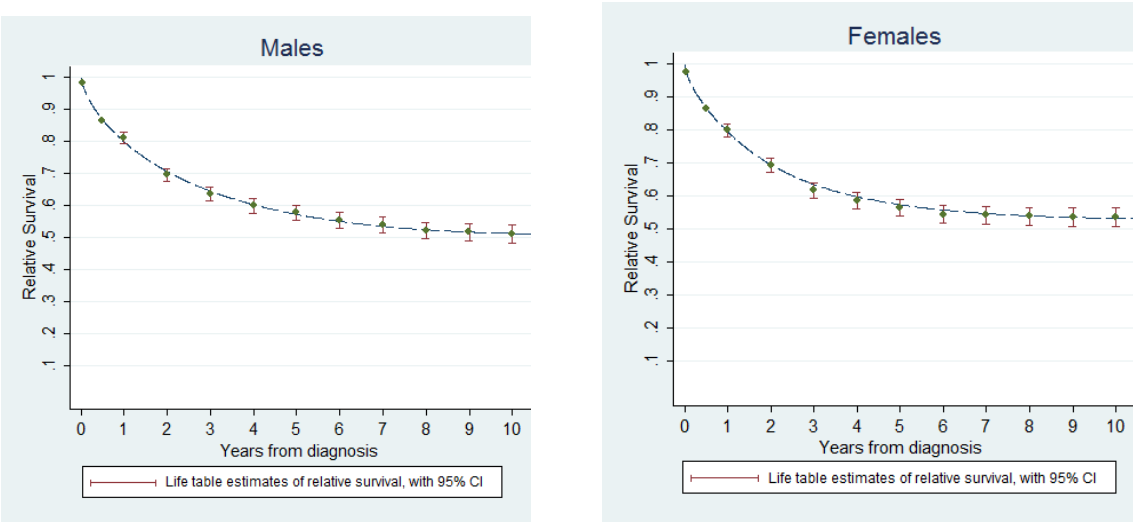

Figure 5.6: Predicted Survival from the Flexible parametric cure model, compared to life table estimates of relative survival for both genders

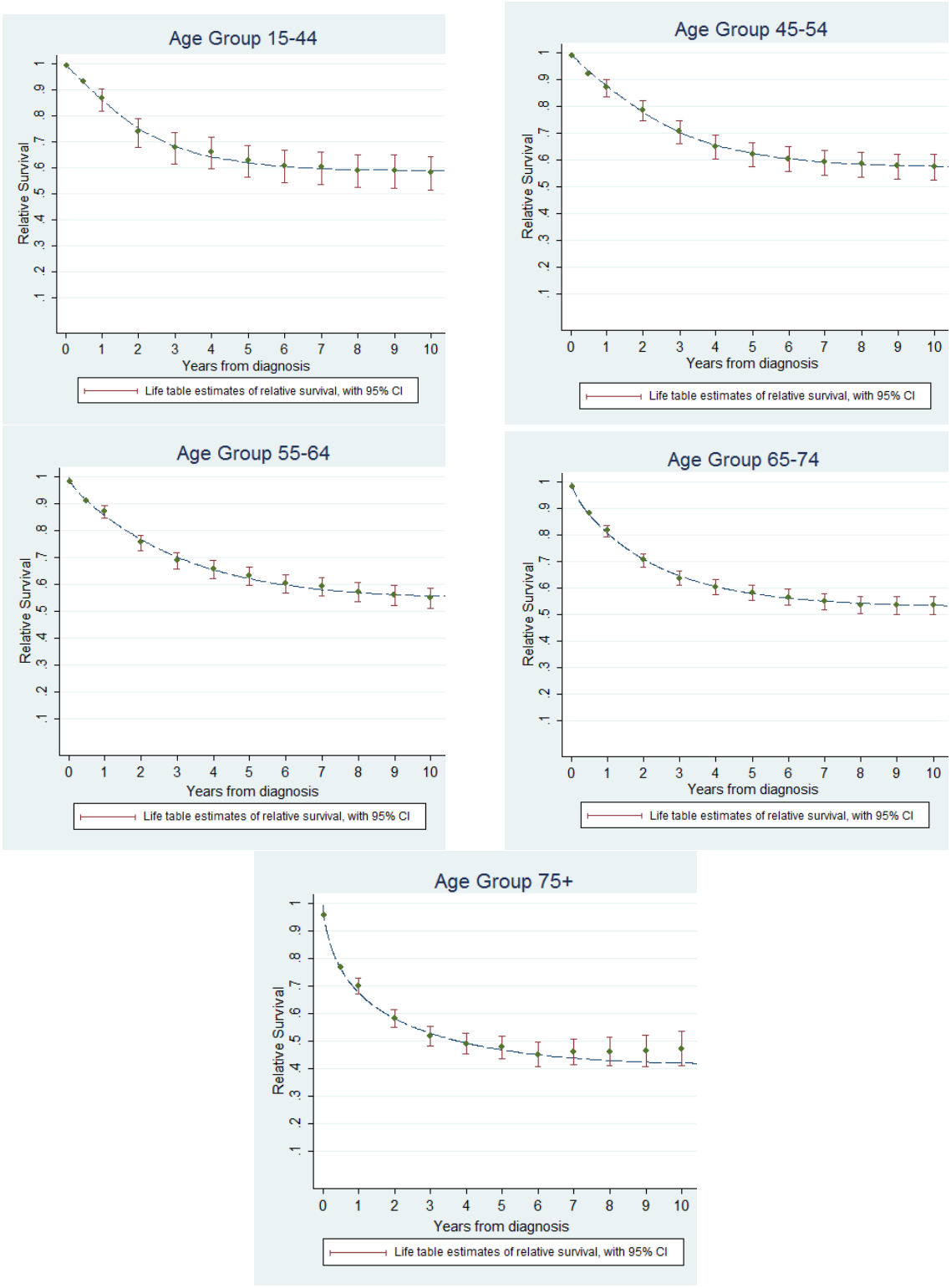

Figure 5.7: Predicted Survival from the Flexible parametric cure model, compared to life table estimates of relative survival for age groups

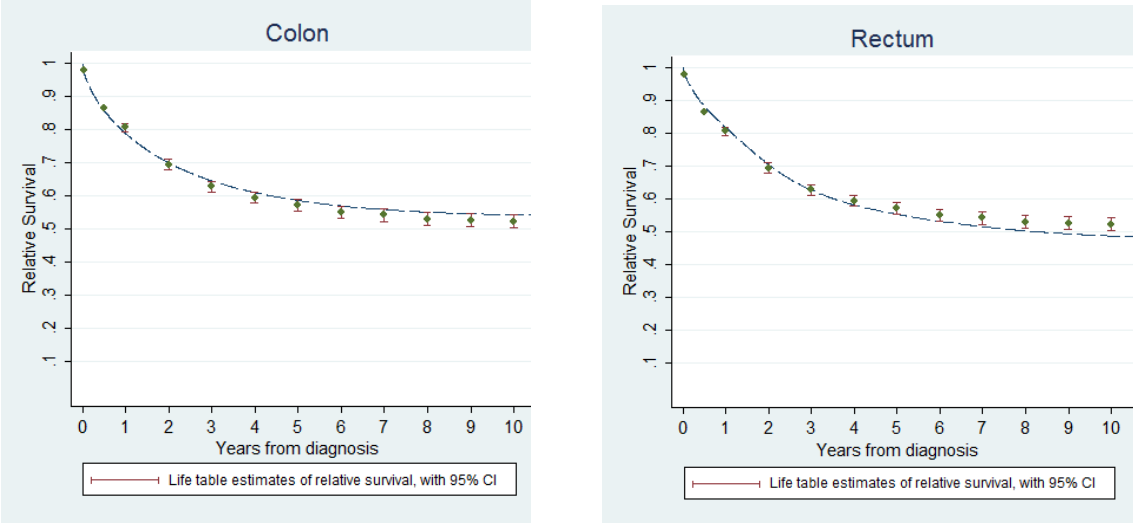

Figure 5.8: Predicted Survival from the Flexible parametric cure model, compared to life table estimates of relative survival for typology

Table 5.7: Estimated proportion cured (%) and 95% confidence intervals by sex and age group

Cancer of the colon

	Age <45	Age 45-54	Age 55-64	Age 65-74	**All <75**
Men	55.13]42.45 , 66.11[	58.43]49.39 , 66.42[	54.93]48.40 , 60.98[	53.86]47.67 , 59.64[	**54.44]50.72 , 58.00[**
Women	66.84]55.21 , 76.08[	61.75]52.70 , 69.58[	58.78]51.66 , 65.21[	58.62]52.53 , 64.21[	**59.68]55.90 , 63.25[**
All	**61.47]52.91 , 68.93[**	**60.01]53.72 , 65.72[**	**56.70]51.88 , 61.22[**	**56.24]51.93 , 60.32[**	

Table 5.8: Estimated proportion cured (%) and 95% confidence intervals by sex and age group

Cancer of the rectum

	Age <45	Age 45-54	Age 55-64	Age 65-74	**All <75**
Men	55.80]39.67 , 69.20[	52.93]42.75 , 62.10[	52.99]44.75 , 60.56[	47.17]39.72 , 54.25[	**50.78]46.17 , 55.20[**
Women	57.66]41.84 , 70.61[	58.24]45.65 , 68.89[	54.87]45.16 , 63.55[	49.00]40.68 , 56.80[	**52.66]47.47 , 57.58[**
All	**56.08]44.93 , 65.82[**	**54.24]46.46 , 61.37[**	**53.43]47.19 , 59.26[**	**48.16]42.62 , 53.47[**	

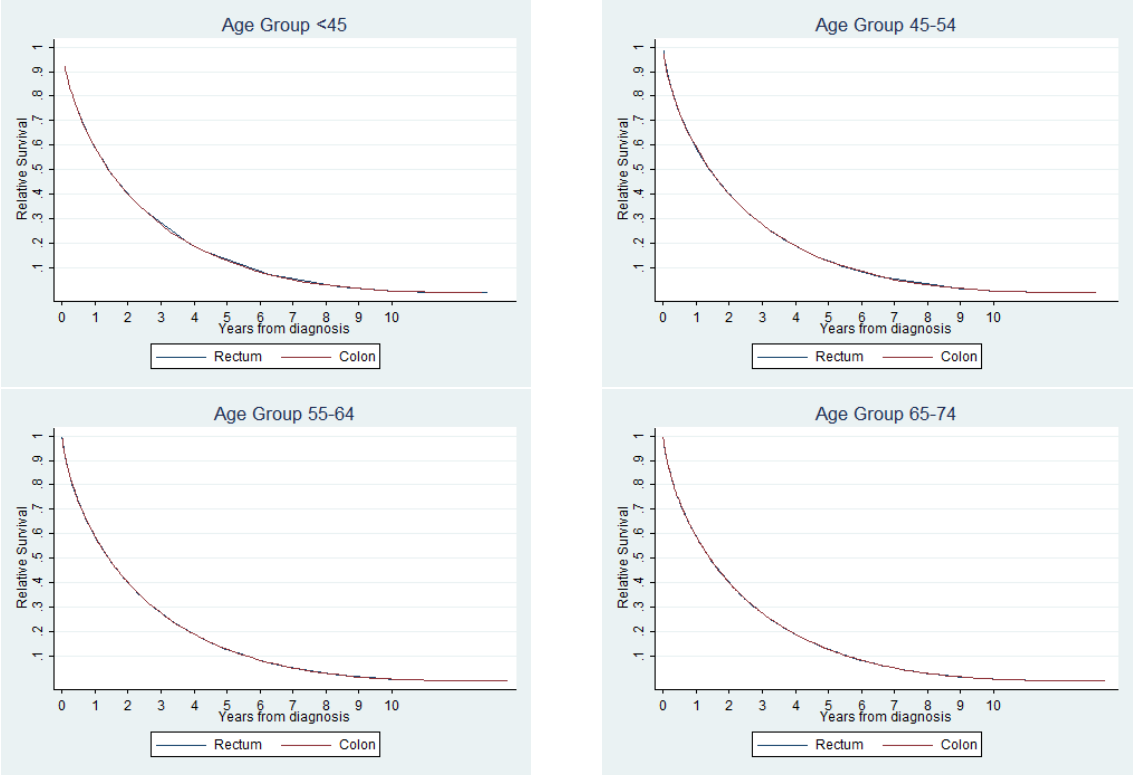

Figure 5.9: Predicted Survival of Uncured for Men

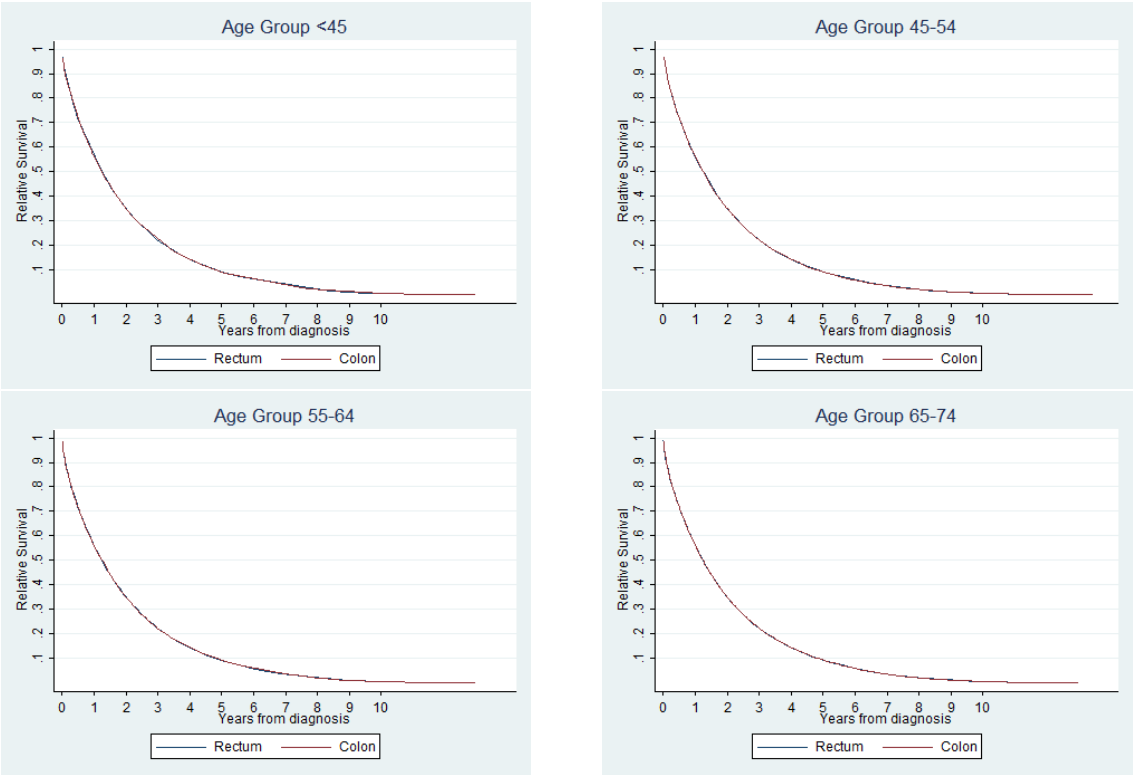

Figure 5.10: Predicted Survival of Uncured for Women

Tables 5.7 and 5.8 summarize the estimated cure percentages results of the final model by topography. As expected, independent on the carcinogen location, the chance of cure decreased with advanced age at diagnosis for both genders. Looking at the results by gender there is evidence in all ages that men are bellow the chances of women beating the cancer. Besides that, the part of the large intestine where cancer develops, whether colon or rectum also seems to influence the cure proportion. Patients in the study with cancer of the rectum revealed lower chances of achieving cure when compared with ones with cancer developed in the colon.

A more detailed analysis on the results of cancer of the colon indicates a percentage of cure above 50% for all ages and genders. Younger female patients up to 45 years old showed a cure percentage of 67%, while in men the cure was 55% with a confidence interval containing percentage lower than 50%. Curiosly, men in the following age group (45-54) revealed better odds with 58%, while women dropped to 62%. Nevertheless, the deviation in the cure proportion compared to women became lower. The two older age groups in the model have similar results for cure proportion, men's ranging from 55% to 54% and women's estimate of about 58%.

5.5 Discussion

In this study, methods of relative survival were attempted to be applied and analyzed to the available data set. Like in the Melanoma case study, specific cause of death was not available for all patients in the data and it was preferable to use relative survival approach instead of cause-specific survival analysis. In section 5.2, survival estimates suggest similar prognosis for both men and women. With a 5-year relative survival rate of 56% and 55%, respectively. Survival curve seems to reach a plateu indicating that patients experience long-term survival. However, when fitting a Cox regression model, statistically significant differences in prognosis between genders were observed. Men, actually, have higher risk of death. Patient age at diagnosed seem to have influence in prognosis, during the follow-up period relative survival estimates are lower as we go from the newest to the oldest age group. Looking at relative survival estimates there also seem to be a distinct correlation between survival and region type in which patients live, with worse prognosis scenarios for ones who came from rural areas. Nevertheless, cox regression model proved no statistical evidence of differences. topography of the cancer, whether it developed in the rectum or in the colon, led to discrepancies in patient's prognosis with statistical significance tested in Cox regression model. Rectum cancer has poorer survival rates and increases in approximately 10% the mortality rate. Colon and rectal cancer

are often referred as colorectal cancer since they share many features. Notwithstanding, many studies have reported differences in survival and cure with advantages in survival for colon cancer patients [36]. Our findings are consistent with previous literature and moreover, is relevant to notice that relative survival estimates, Cox regression model and cure model approach are in accordance and can be seen as complementary methods. Relative survival estimates provide an overview to the survival over time since diagnosis, hazard ratios from Cox model evaluate which features are more associated with success results on survival and cure model indicates the cure percentages and the prognosis for those who are not able to achieve cure. Figure 5.9 and 5.10 shows that patients who are not likely to achieve cure have survival tending to zero, as expected since they actually die. Their survival dramatically decreases within the first 5 years of follow-up to approximately 0% for all ages, genders and even tumour topography. It would be interesting if the available data could provide records over a longer period of time, with registrations before 1998. In that case we would be able to study temporal trends in survival, cure and also perceive the underlying reasons for those trends. Thus, future population-based data studies to be performed will play an important role in determining whether or not there is a meaningful progress in patient survival and cure achievement. Possibility of cure for colorectal cancer is reasonable and we may roughly say it is above 50%.

Chapter 6

Conclusion

In this thesis, methods for assessment and modelling of cancer survival were presented. It was given special attention to a new class of survival regression models to estimate cure fraction. The main aim was to apply this new models and demonstrate their value to survival analysis as a complement to the standard techniques. Population-based data from colorectal and melanoma skin cancer was provided from North Region of Portugal Cancer Registry to perform the study. Relative survival rates and Cox regression model provide an overview to the survival during the follow-up period and the potential prognostic risk factors. However, information on cure percentages is only available when applying cure models. Since this method is only reliable when cure is a reasonable assumption, it enables to distinguish patients into ones who will achieve cure and ones who do not. Thus, besides estimating cure proportion it also give information on the prognosis for the uncured group. This is a great advantage of this approach, because it allows to divided patients into two different groups with separete survival distributions for each.

Our findings can be useful in understanding cancer survival for colorectal and melanoma skin cancer in the Northern Portugal. We found that, for both cancers in the study, men have higher risk of dying during the follow-up compared to women. Survival rates and cure percentages are actually lower for patients diagnosed with colorectal cancer in comparison with patients with melanoma skin cancer. We can roughly say that a patient diagnosed with melanoma has a probability of beating the disease above 65%. However, if diagnosed with colorectal cancer the probability decreases. Besides sex and age factors, anatomical site of the cancer play an important role in the prognosis, with patients whose cancer has developed in the rectum having worse survival.

Survival and cure patterns reflect the underlying response to cancer treatment and suggest trends over time. Results can elucidate if the adopted health care techniques are contributing to long-term survival or even better to reaching cure. Future researches in

combination with these findings will potentially provide insights into how treatment and prevention methods advances are changing survival and ultimately impacting cure.

References

[1] Lindsey a Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-tieulent, and Ahmedin Jemal. Global Cancer Statistics, 2012. 00(00):1–22, 2015.

[2] Direção de Serviços de Informação e Análise Programa Nacional para as Doenças Oncológicas. Portugal, doenças oncológicas em números. 2013.

[3] Population-based Public Health Practice. Population-Based Public Health Practice. pages 3–16, 2010.

[4] Jorge N. Izquierdo and Victor J. Schoenbach. The potential and limitations of data from population-based state cancer registries. *American Journal of Public Health*, 90(5):695–698, 2000.

[5] Washington County. Population-based Cohort Studies. 20(1):81–90, 1998.

[6] Hyunsoon Cho, Nadia Howlader, Angela B Mariotto, and Kathleen A Cronin. Estimating relative survival for cancer patients from the SEER Program using expected rates based on Ederer I versus Ederer II method. 2011.

[7] Maja Pohar Perme, Janez Stare, and Jacques Estève. On estimation in relative survival. *Biometrics*, 68(1):113–20, March 2012.

[8] Diana Sarfati, Tony Blakely, and Neil Pearce. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International journal of epidemiology*, 39(2):598–610, April 2010.

[9] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:187–220, 1972.

[10] Patricia M. Grambsch Terry M. Therneau. *Modeling Survival Data: Extending the Cox Model*. 2000.

[11] H G Welch, L M Schwartz, and S Woloshin. Are increasing 5-year survival rates evidence of success against cancer? *JAMA : the journal of the American Medical Association*, 283(22):2975–8, June 2000.

[12] Patrick Royston. Flexible parametric alternatives to the Cox model , and more. *Stata Journal*, 1(1):1–28, 2001.

[13] Dickman PW. Lambert PC1, Thompson JR, Weston CL. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 2007.

[14] P.W. Dickman X.Q. Yu, R. De Angelis, T.M.L. Andersson, P.C. Lambert, D.L. O'Connell. Estimating the proportion cured of cancer: Some practical advice for users. *Cancer Epidemiology*, 37(6):836–842, 2013.

[15] a Verdecchia, R De Angelis, R Capocaccia, M Sant, a Micheli, and G Gatta. The cure for colon cancer:Results from the eurocare study. *Int.J.Cancer*, 77(February):322–329, 1998.

[16] Paul C. Lambert, Paul W. Dickman, Pia Österlund, Therese Andersson, Risto Sankila, and Bengt Glimelius. Temporal trends in the proportion cured for cancer of the colon and rectum: A population-based study using data from the Finnish Cancer Registry. *International Journal of Cancer*, 121(9):2052–2059, 2007.

[17] Jeremy M.G Taylor Fabien Corbière, Daniel Commenges and Pierre Joly. A penalized likelihood approach for mixture cure models. *Statistics in Medicine*, 28(3):510–524, 2009.

[18] B. Soderman R. De Angelis, R. Capocaccia, T. Hakulinen and A. Verdecchia. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18(4):441–454, 1999.

[19] Kathleen A. Cronin Binbing Yu, Ram C. Tiwari and Eric J. Feuer. Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine*, 23(11):1733–1747, 2004.

[20] Therese M L Andersson and Paul C. Lambert. Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models. *Stata Journal*, 12(4):623–638, 2012.

[21] Paul C. Lambert. Modeling of the cure fraction in survival studies. *Stata Journal*, 7(3):351–375, 2007.

[22] Michael J. Crowther & Paul C. Lambert Mark J. Rutherford. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85(4):777–793, 2013.

[23] Sylvain Durrleman and Richard Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.

[24] D. R. Cox. Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, 83(404):187–220, 1988.

[25] M B Lens and M Dawes. Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma. *The British journal of dermatology*, 150(2):179–85, February 2004.

[26] J. Galceran E. Ardanaz R. Schaffar A. Pring M. Arnold, C. Holterhues, L.M. Hollestein, J.W.W. Coebergh, T. Nijsten, E. Pukkala, B. Holleczek, L. Tryggvadóttir, H. Comber, M.J. Bento, Ch.S. Diba, R. Micallef, M. Primic-Žakelj, M.I. Izarzugaza, J. Perucha, R. Marcos-Gragera and E. de Vries. Trends in incidence and predictions of cutaneous melanoma across Europe up to 2015. *Journal of the European Academy of Dermatology and Venereology*, 28(9):1170–1178, 2014.

[27] By Charles M Balch, Seng-jaw Soong, Jeffrey E Gershenwald, John F Thompson, Douglas S Reintgen, Natale Cascinelli, Marshall Urist, Kelly M Mcmasters, Merrick I Ross, John M Kirkwood, Michael B Atkins, John A Thompson, Daniel G Coit, David Byrd, Renee Desmond, Yuting Zhang, Ping-yu Liu, Gary H Lyman, and Aberto Morabito. Prognostic Factors Analysis of 17 , 600 Melanoma Patients : Melanoma Staging System. 19(16):3622–3634, 2001.

[28] Jade Homsi, Mohammed Kashani-sabet, Jane L Messina, and Adil Daud. Cutaneous Melanoma : Prognostic Factors. pages 223–229, 2005.

[29] Divisão administrativa. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_cont_inst&INST=6251013&xlang=pt, 2014.

[30] Timo Hakulinen and Kamal H. Abeywickrama. A computer program package for relative survival analysis. *Computer Programs in Biomedicine*, 19(2-3):197–207, January 1985.

[31] Reintgen DS Austin PF, Cruse CW, Lyman G, Schroer K, Glass F. Age as a prognostic factor in the malignant melanoma population. *Ann Surg Oncol.*, (6):487–94, 1994.

[32] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W W Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6):1374–1403, 2013.

[33] Ahmedin Jemal, Robert a Smith, and Elizabeth Ward. Worldwide Variations in Colorectal Cancer. 2009.

[34] Rachet B. Shack LG, Shah A, Lambert PC. Cure by age and stage at diagnosis for colorectal cancer patients in North West England, 1997–2004: A population-based study. *Cancer Epidemiol*, 36(6):548–553, 2012.

[35] Registo Oncologico Regional Centro. Registo Oncológico Nacional 2008. 2014.

[36] Gu J. Li M, Li JY, Zhao AL. Colorectal cancer or colon and rectal cancer? Clinicopathological comparison between colonic and rectal carcinomas. *Oncology*, 73, 2007.